REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author WHITE D.I

This thesis comes within category D.

☑ This copy has been deposited in the Library of UCL

☐ This copy has been deposited in the Senate House Library, Senate House, Malet Street, London WC1E 7HU.

EVOLUTIONARY HISTORY OF A SOUTH AMERICAN POPULATION ISOLATE AND THE GENETIC BASIS OF A COMPLEX NEUROPSYCHIATRIC TRAIT

by

Daniel James White

A Thesis Submitted for the Degree of Doctor of Philosophy at the University of London

The Galton Laboratory
Department of Biology
University College London

September 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Much has been learnt about the genetics of *Homo sapiens* over the last 130 years including gene structure and number, genome size, and levels of genetic diversity. Many things remain less well understood, however, not least the genetic basis of common, complex traits and disorders. Consideration of functionally important but non-coding regions may improve understanding. I assessed naturally occurring, genetic variation in the promoters of four serotinergic genes and revealed (75%) to be polymorphic, two-thirds of which (50% overall) had functional variant haplotypes. These promoter-based polymorphisms are good functional candidate loci for psychiatric trait mapping studies. The variation within our genomes is organised by historic evolutionary and demographic events. Populations with unique demographic histories may be important in complex trait gene mapping, and the Antioquia isolate (North-West Colombia) is an example of such a population. Using population genetic analyses I have shown high autosomal diversity in Antioquia; structure analysis showed relatedness to be strong with Spain and modest with African and Native American populations, likely reflective of its historic admixture. LD was not pronounced in Antioquia, potentially an artefact of marker selection. However, Antioquia may have an important role in admixture mapping. Mapping multifactorial psychiatric traits and disorders is particularly challenging for geneticists. To investigate the genetics of BPI, I performed a family-based association analysis of the *SLC6A4* gene in the Antioquia and CVCR populations using 10 SNPs, 3 STRs and 1 VNTR spanning approximately 300kb, including an assessment of LD structure. Moderate over-transmission in BPI cases was observed for a haplotype consisting of the functional VNTR (the LPR) long allele and an adjacent STR (Antioquia TDT χ^2= 6.00, p=0.014; CVCR HHRR χ^2=5.012, p=0.025; both TDT χ^2= 8.00, p=0.005). Characterising genetic variation at the population level is important to improve population-based genetic association studies of complex traits, and the inclusion of regulatory variation is supported.

Key Abbreviations; LD: linkage disequilibrium, BPI: bipolar affective disorder class 1, LPR: length polymorphic region, CVCR: Central Valley of Costa Rica.

Dedication

In memory of my grandmother, Kathleen Scott, who sadly passed away while I researched human genetic variation at the University of California, Los Angeles.

Acknowledgements

For help in completing this large chapter of life there are many people that I would like to acknowledge. In the U.K. first and foremost thanks must go to my supervisor, Dr. Andres Ruiz-Linares, whose valuable time and knowledge has always been appreciated, as were the thoughtful suggestions of my secondary supervisor, Professor Ziheng Yang. Colleagues from the group that have given me invaluable support include, in particular, Nicolas Trujillo-Pineda and Dr. Luis Carvajal-Carmona, while Ibi Herzberg-Wallbank, Sijia Wang and Barbara Kremeyer have all made life easier, and Lupe Carvajal-Carmona's assistance in the lab was always appreciated. From slightly further afield, interactions with Dr. Claire Willoughby, Steve Jeremiah and Professor Dallas Swallow's group were always useful. Thanks also go to Ian Evans and Dr. Jacques Gianino. A special mention must go to Professor Sue Povey and Luis for helpful comments on this dissertation.

From the U.S. mentorship of the highest calibre came from Dr. Charles Glatt, problems became surmountable while working with Dr. Michael DuPree and no data would have been collected without assistance from Margaret Chu, Thuy Vu and Maricel Tampicilic. All of this under the calm leadership of Dr. Nelson Freimer, and wise words from Drs. Sue Service and Roel Ophoff. No less valued were the friendships of Anna Grygoruk, Sherry Breidenthal and Estreya Kapuya.

Outside the scope of this thesis several people need thanked for keeping me relatively sane. These include Luke Warren and Gianpiero Cavalleri for important on-site respite, Pete Little for important off-site respite, Melanie Scott for laughing at my jokes, Lindsay and Simon Warner for family support of the best kind and my mother for always caring.

This work was made possible by a research grant from the Biotechnology and Biological Sciences Research Council.

Contents

List of Tables

List of Figures

List of Abbreviations

Ant	Antioquia
AMH	Anatomically Modern Human
BPAD, BP	Bipolar Affective Disorder
BPI	Bipolar Affective Disorder, type 1
CD/CV	Common Disease / Common Variant hypothesis
Chip	Chipewayan
cSNPs	Coding SNPs
CVCR	Central Valley of Costa Rica
dsDNA	Double stranded DNA
D	Linkage Disequilibrium Coefficient
D'	D/Dmax
Dmax	Maximum value of D given allele frequencies
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, 4th edition
FET	Fisher's Exact Test
FETp	Fisher's Exact Test probability value
HGNC	HUGO Gene Nomenclature Committee
HGP	The Human Genome Project
HHRR	Haplotype-based Haplotype Relative Risk
HUGO	The Human Genome Organisation
Kb	Kilobases
LD	Linkage Disequilibrium
MAF	Minor allele frequency
Mb	Megabases
ML	Maximum likelihood
NCBI	National Center for Biotechnology Information
n	Sample size
Ne	Effective population size
ng	Nanograms
p	Frequency
p	Probability
PCR	Polymerase Chain Reaction
R	Recombination fraction
pmol	Picomole
sd	Standard deviation
se	Standard error
SSCP	Single strand conformation polymorphism
STR	Short tandem repeat
TBE	Tris-borate EDTA
TDT	Transmission disequilibrium test
UCLA	University College London
UTR	Untranslated region
V	Volts
VNTR	Variable number of tandem repeats
5HTTLPR, LPR	Serotonin transporter length polymorphic region
θ	Population mutation parameter
μl	Microlitres
π	Nucleotide diversity

CHAPTER 1: INTRODUCTION

CHAPTER 1: INTRODUCTION

1.1 Genetics of *Homo sapiens*

1.1.1 The Eukaryotic Genome

Every living organism has a particular arrangement of genes in their genome that forms a plan for its growth and development. Genomes are composed from four nitrogenous bases - adenine and guanine (the purines), and cytosine and thymine (the pyrimidines) - which, along with a sugar molecule and phosphate molecule, form the nucleotide building blocks of the genetic code. These nucleotides are specifically arranged to make deoxyribonucleic acid (DNA), most stable as a double stranded helix (dsDNA). In eukaryotes genomic dsDNA is packed together with chromatin and histone proteins as smaller units, the chromosomes, and stored in cellular nuclei. Eukaryotes can be haploid and carry one copy of a genome, or diploid in which there are two copies, one passed from each parent.

There is a lot of variation across eukaryotes in genome size and structure. For example, genomes can consist of around 10Mb (megabasepairs) to 200,000Mb of dsDNA, and may contain from 2 to over 30 thousand genes (Koonin, 2003). In general, a rough trend that is followed is as the number of vital processes that occur within an organism increase, often regarded as an organism's complexity, so the number of protein coding genes required increases. Equally, a strong correlation between the number of genes in an organism's genome and the DNA quantity may be expected. However, this correlation is considerably less apparent in eukaryotes than prokaryotes (Koonin, 2003). For example, some insects may have a haploid DNA content (C factor) of 100Mb, compared to some unicellular amoeba with a C factor of 686,000Mb (Li, 1997); some 7,000 times more DNA than the more complex, developed group (although significantly less protein coding genes). This phenomenon is referred to as the C-value paradox and is still not fully understood (Petrov, 2001).

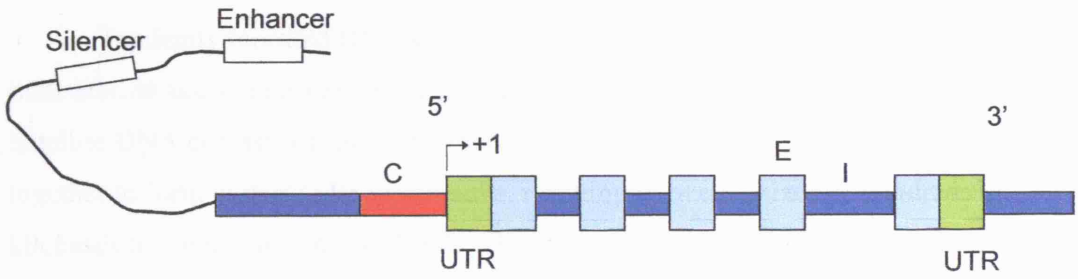
1.1.1.1 Non-coding DNA in the eukaryotic genome

It is apparent, therefore, that it is not the number of protein coding genes that account for the dramatic variation in eukaryotic genome size; instead this is due to variable genomic levels of non-coding DNA. In fact, the majority of DNA in the

genomes of more complex eukaryotes is non-coding (Li, 1997). For example, in mammals no more than 3% is coding; the remainder is non-coding, some functional but most non-functional.

Non-coding DNA can either be intragenic (located within genes) or intergenic (located elsewhere in the genome between genes). Eukaryotic genes are not made solely from coding DNA as many genes have a complex arrangement of coding exons separated by non-coding introns and bordered by 5' and 3' untranslated regions (UTRs). In addition, important promoter sequence responsible for the regulation of gene expression is located upstream (5') to the first exon (see figure 1.1).

Figure 1.1. Components of a Eukaryotic Gene



The major components of a eukaryotic gene are shown. Only the exons represent coding DNA. The core promoter often contains transcription factor binding sites that are recognised by component proteins of the complex that drives expression. C: core promoter; UTR: untranslated region; E: exon; I: intron; + 1: transcription initiation site from where the messenger ribonucleic acid (mRNA) molecule is assembled.

1.1.1.2 Repetitive DNA in the eukaryotic genome

Although some non-coding DNA is intragenic, the vast majority resides in stretches of chromosome that separate genes. A large proportion of this exists as repetitive sequence, with estimates ranging from 20% in yeast to 90% in arthropods (Li, 1997). Repetitive DNA can consist of; (i) repeat units dispersed across the whole genome such as LINEs (long interspersed elements) and SINEs (short interspersed elements), or (ii) local, tandemly repeated sequence.

LINEs are abundant, self-proliferating, repetitive DNA sequences at least 5kb in length and can have up to 100,000 copies in mammalian genomes (Esnault et al., 2000). Early work showed that they are derived from POL II transcripts (Deninger,

1989; Hutchison et al., 1989), and part of their sequence encodes reverse transcriptase, used in self-proliferation. Hence, these elements are often referred to as retroposons or retrosequences. SINEs are shorter than LINEs at 75-500 bp in length, do not use reverse transcription for self-proliferation but are able to retrotranspose and are regarded as non-functional retropseudogenes (Dewannieux et al., 2003). They are derived from POL III transcripts (Deninger, 1989; Hutchison et al., 1989) and are particularly abundant in mammalian genomes. Many SINEs are described as being *Alu* or *Alu*-like as it was this SINE family that was characterised first (Schmid and Shen, 1985). The *Alu* family derives its name from possession of an *AluI* restriction endonuclease site, and in primates is composed of a dimer of two similar sequences believed to have evolved from the 7SL RNA gene (Strachan, 2003; Ullu et al., 1982; Ullu and Tschudi, 1984).

Tandemly repeated DNA sequence comes in 3 major classes - satellite, minisatellite and microsatellite – and repeat arrays are found throughout the genome. Satellite DNA consists of low order repeat units from 2 to 100s of bps, which can join together to form higher order repeat units, resulting in overall sizes of hundreds of kilobases to megabases (Avise, 2004). They are commonly found in heterochromatic DNA near centromeric regions, and a role in ensuring the centromere is the last chromosomal region to replicate during mitosis has been suggested (Csink and Henikoff, 1998). Satellites can form a large part of eukaryotic genomes, and up to 50% in some mammals (Brown, 2002; Comings and Okada, 1976; Hatch et al., 1976; Strachan, 2003). Minisatellites are composed of repeat units of 7-100 bps, repeated from 10 to 100 times, and are not confined to particular genomic regions. The third major class of tandemly repeated DNA are short tandem repeats (STRs) or microsatellite DNA. Microsatellites have repeat units from 1-6 bp and may also be repeated from 10 to 100 times. VNTR (variable number of tandem repeats) is a collective description for mini- and microsatellites, although more commonly reserved for minisatellites. Characteristically, VNTRs have high mutation rates, are widely distributed throughout the genome and are of a size that enables easy amplification by the polymerase chain reaction (PCR). This makes them very useful tools in genetic analyses such as paternity and maternity testing, generation of genetic maps and associating genes with traits and diseases (Avise, 2004).

Table 1.1 Repetitive DNA in the eukaryotic genome

Type	Size of Repeat Unit (bp)	Copy Number	How Arrayed
LINE	≥ 5000	100s to 100,000s	Dispersed
SINE	75 to 500	100s to 100,000s	Dispersed
Satellite	2 to 100s (low order) 1000 to 10,000 (high order)	A few to 1000s 100s	Tandem
Minisatellite	7 to 100	10 to 100	Tandem
Microsatellite	1 to 6	10 to 100	Tandem

LINE: long interspersed elements; SINE: short interspersed elements.

1.1.2 The Human Genome

Our DNA is packed into twenty-three chromosomes; twenty-two autosomes and either an X or Y sex chromosome. As we are diploid there are two copies of each autosome in our somatic cells, and either two X chromosomes or an X and a Y, depending if female or male respectively. Additionally, human mitochondrial DNA (mtDNA) forms part of the human genome and is important in many of our cellular processes. It is unique in that it is clonally inherited through the maternal line in the oocyte cytoplasm.

The most recent version of the human genome sequence at the National Center for Biotechnology Information (NCBI), build 35 version 1 (4th June, 2005), contains 3.02×10^9 basepairs in the haploid human genome, of which 2.83×10^9 are confirmed by more than one source and therefore considered reliable (http://www.ncbi.nlm.nih.gov/). Current estimates place the approximate total of nucleotides in the genome at 3.08×10^9 bps (Collins et al., 2004). The mtDNA has been fully sequenced and contains 16 571 basepairs representing approximately 0.0005% of the human genome. Interestingly, it contains 37 intronless genes, making 93% of its sequence coding (Strachan, 2003).

1.1.2.1 The Human Genome Project

The majority of human DNA sequence has been made publicly available by the Human Genome Project (HGP) (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml). The HGP is a collaborative project run by the International Human Genome Sequencing

Consortium (IHGSC) and initially comprised of sequencing centres in the U.S.A., funded to a large extent by the National Human Genome Research Institute, and U.K., organised by the Wellcome Trust Sanger Institute. The project was founded in the mid 1980's by the US Department of Energy who were interested in obtaining knowledge of the effect of variations in the human genome on medical health. Later additions to the IHGSC came from France, Germany, China and Japan. To help maintain communication and organisation between all centres the Human Genome Organisation (HUGO) group was set up in 1988. Interest, funding and organisation gained momentum resulting in a published draft genome in February 2001 (Lander et al., 2001). In April 2003 the project was declared essentially complete as more than 98% of euchromatic DNA had been sequenced by twenty sequencing centres in six different countries (Collins et al., 2003). The next major challenge for geneticists is the accurate annotation of the genome.

1.1.2.2 Organisation of the Human Genome

How its various components are organised is the next important stage in the understanding of our genome. Recently, considerable effort has been made to accurately annotate the human genome, and work has proceeded on a chromosomal basis. A summary for six chromosomes is presented in table 1.2.

Table 1.2. Annotation summary of 6 human chromosomes.

	Size (x 10^6 bp)[1]	No. Genes[2]	% Non-Coding	% Repetitive
Chrom 2[a]	~ 237	1,346	91.4	46.1
Chrom 4[a]	~ 186	796	95.8	48.6
Chrom 5[b]	177.7	923	na	46.3
Chrom 9[c]	109.0	1,149	97.5	46.2
Chrom 10[d]	131.7	1,357	97.7	43.7
Average (5 autosomes above)	**168.3**	**1114.2**	**95.6**	**46.2**
X chrom[e]	151.0	1,098	98.3	56[3]

[1]Inclusive of 99.1 to 99.7% of euchromatic sequence. [2]Genes refer to known and predicted protein-coding genes, pseudogenes are not included. [3]Inclusive of interspersed repeats only. na: not available. [a](Hillier et al., 2005), [b](Schmutz et al., 2004), [c](Humphray et al., 2004), [d](Deloukas et al., 2004), [e](Ross et al., 2005).

Determining the exact number of genes in the human genome has not yet been achieved and awaits further improvement of gene detection techniques. Early estimates of total gene number were between 65,000 and 80,000 (Antequera and Bird,

1993; Fields et al., 1994). Antequara and Bird (1993) used the numbers of CpG islands in the genome to estimate gene number by making several assumptions such as CpG islands are always associated with genes and that 56% of all genes have a CpG island. Other estimates are based on the number of unique expressed sequence tags (ESTs) and sequencing analysis of sub-sections of the genome (Fields et al., 1994). However, sequence made available by the HGP provides different estimates. In July 2004 NCBI recorded only 15,630 protein coding genes with known or inferred function, 125 ribosomal genes that encode for the transcription and translation machinery and a further 5,832 protein coding genes with no known function (http://www.ncbi.nlm.nih.gov/LocusLink/statistics.html). Currently (July 2005) the HUGO Gene Nomenclature Committee (Wain et al., 2002) supports 21,697 approved gene symbols in the Genew database (http://www.gene.ucl.ac.uk/nomenclature/). This evidence suggests that there are less genes in the human genome than previously thought and that other mechanisms may be responsible for the phenotypic complexity of *Homo sapiens* and other higher vertebrates.

To determine how much of the human genome is coding estimations can be made from available data. If 25,000 genes are assumed to exist in the genome (Collins et al., 2004) and they have an average size of 27 kb (Lander et al., 2001) then approximately 23% of the genome is composed of genic DNA. If the average protein is 480 amino acids (Lander et al., 2001), corresponding to 480 codons and 1340 bps of DNA, then approximately 1.1% of the genome is coding. This figure will likely increase as gene detection methods are improved, made possible by the completed genomes of not only humans but many other species. Nonetheless, this estimate is consistent with others (Collins et al., 2004) and implies that at least 98% of the genome is non-coding. Although the majority may be redundant, understanding of the functional relevance of the non-coding genome is gradually improving.

1.1.2.3 The Non-Coding Human Genome

Extragenic sequence in humans, like other eukaryotes, is composed of several types of DNA including single copy, rarely repeated DNA as well as a large amount of relatively highly repeated sequence. Around 45% of the genome is reported to consist of repetitive interspersed DNA (Deloukas et al., 2004; Humphray et al., 2004; Lander et al., 2001; Ross et al., 2005; Schmutz et al., 2004). Estimates suggest 21% can be accounted for by LINEs (the L1 family being the most common, for example

29% of the X chromosome consists of L1), 13% can be accounted for by SINEs (there are 1,558,000 copies in the human genome of which 1,090,000 are *Alus*), 8% by retrovirus-like elements and 3% by transposon copies (Dunham et al., 1999).

There is also a large quantity of tandemly repeated DNA. For example, alpha satellite DNA is a common component of human chromosomal centromeres. Additionally, approximately 40% of the heterochromatic part of the non-recombining Y chromosome is reportedly comprised of one satellite, DYZ1 (Ali and Hasnain, 2002; Nakahori et al., 1986). Minisatellites are found distributed throughout the genome, although the common GC class of minisatellites are often clustered towards the ends of chromosomes (Jeffreys et al., 1998; Royle et al., 1988). Microsatellites are more evenly distributed throughout the genome and, like all satellite DNA, are very common in the human genome (Subramanian et al., 2003). For example, the genome contains around 35,400 trinucleotide and 97,500 tetranucleotide microsatellites (Lander et al., 2001). The most common dinucleotide is the CA repeat and may represent 0.25% of total genomic DNA (Lander et al., 2001).

1.1.3 Variation in the Human Genome

There is a substantial amount of variation between the genomes of individuals, and may reside in both coding and non-coding regions. Although some variation will have been selectively advantageous and been selected for in the past, most newly arisen, functional genetic variation (for example mutations in coding DNA at non-synonymous sites) is not expected to persist in populations; it is likely to have a detrimental effect and so be removed by purifying selection (Cargill et al., 1999). Conversely, if mutations arise in non-functional DNA the actions of purifying selection may be evaded as this variation will be selectively neutral. Kimura's theory of neutral evolution, proposed in 1968, describes how selectively neutral variation is generated by mutation and lost by genetic drift, and at equilibrium a balance between these two factors can maintain substantial levels of polymorphism within populations (Hartl and Clark, 1997; Kimura, 1968). Further, estimates of mutation rates in non-functional sequence are higher than for the point mutation genome-wide average, which are in the order of 10^{-8} per basepair per generation (Arndt et al., 2005; Kumar and Subramanian, 2002; Nachman and Crowell, 2000). Consequently, substantial selectively neutral, genetic variation may exist between individuals and populations.

Two important sources of variation of particular use to genetic research are microsatellites and single nucleotide polymorphisms (SNPs).

1.1.3.1 Microsatellites

As mentioned previously microsatellites are common features in the human genome and are found both intergenically and intragenically. Mutation rates are very high in these loci with estimated rates of 10^{-3} to 10^{-4} per locus per generation, about 10^5 fold greater than nucleotide substitution rates (Brinkmann et al., 1998; Holtkemper et al., 2001; Huang et al., 2002; Xu et al., 2000).

The mutation mechanism of microsatellites likely involves replication slippage, where an array loses or gains a repeat unit during DNA replication (Kayser et al., 2000). Often the stepwise mutation model, in which contractions and expansions occur with equal probability at a fixed rate, can explain microsatellite evolution (Di Rienzo et al., 1998). However, more recent work has shown that mutation in microsatellites may be generated by additional means. For example, although most mutations do involve a change of a single repeat unit, it has become apparent that contraction mutations, where the number of arrays decreases, are more likely when the number of arrays is large (Xu et al., 2000). There is also evidence that mutation rate increases with array length suggesting the stepwise mutation model may be too simple (Brinkmann et al., 1998). Further, replication slippage may not be the only mechanism, as degree of heterozygosity in a population may also increase mutation rate suggesting an inter-allelic effect such as non-homologous recombination or gene conversion (Cooper et al., 1998; Jeffreys et al., 1999; Jeffreys et al., 1998).

Initially microsatellites were considered to be selectively neutral, however more recent work has revealed many may be functional. Roles in transcript splicing (Hui et al., 2003), transcription regulation (Contente et al., 2002) and recombination (Wahls et al., 1990) have all been shown. Furthermore, intragenic trinucleotides are known to have severely detrimental effects. For example, a CAG repeat located in an exon of the HD gene gives rise to Huntington's disease (Macdonald et al., 1993), a promoter based CGG microsatellite leads to Fragile X syndrome (Verheij et al., 1993) and frame shifts generated by microsatellite instability are known to be associated with types of cancer (Kloor et al., 2005). Although the majority of microsatellites may lie intergenically and have no phenotypic effect; the importance of microsatellites to the human phenotype is starting to be realised.

1.1.3.2 SNPs

A very important form of variation at the nucleotide level in the human genome is the single nucleotide polymorphism (SNP), formed when a nucleotide at a given site has been replaced by one of the other 3 nucleotides. The minor allele is often required to be present at a population frequency of at least 1%, to enable detection using reasonable sized sample sizes (Cargill et al., 1999).

SNPs are very abundant in the genome. Build 124 (6/1/05) of the SNP database at NCBI (dbSNP) contains 10,054,521 non-redundant human SNPs, of which 5,054,675 have been verified (http://www.ncbi.nlm.nih.gov/SNP). This figure gives rise to a frequency of 1 verified SNP per 598 bps of human DNA, and if all SNPs are included as many as 1 SNP every 306 bps. Both of these figures are significantly more than initial estimates made before the draft human sequence was complete (Kwok et al., 1996; Miller and Kwok, 2001). These numbers continue to increase with ongoing sequence analysis work, particularly in less well studied non-coding and non-functional regions of the genome.

The number of SNPs in coding regions of DNA (cSNPs) had originally been estimated at around 500,000 or approximately 6 per gene (Collins et al., 1998). Using the more recent data above we can estimate that if 10,000,000 polymorphic nucleotide sites exist spread evenly around the genome, and if each gene in the human genome uses on average1340 bps of coding sequence (Lander et al., 2001), then there would be approximately 4.4 cSNPs per gene.

The above calculation is based on a uniform distribution of neutral SNPs; however variation is not expected to be equally distributed throughout the genome. Point mutations may have substantial biological consequence if they disrupt codons or important regulatory sequence. For example a non-synonymous nucleotide substitution could ultimately result in altered protein structure and function. As such, newly arisen alleles in functional genetic regions are not expected to persist in a population (Cargill et al., 1999; Miller and Kwok, 2001) and this principle extends to functional non-coding regions including the regulatory regions of genes. Investigations into levels of nucleotide diversity within the genome consistently report lower levels at non-synonymous sites than elsewhere (Cargill et al., 1999; Glatt et al., 2001; Glatt et al., 2004).

1.1.3.3 Uses of Microsatellites and SNPs

Several characteristics of microsatellites make them particularly suitable for human genetic research: a high mutation rate; their abundance throughout the genome; and the relative ease with which to locate and genotype them. As such, they have illuminated many fields including DNA fingerprinting for forensic and paternity studies (Kayser and Sajantila, 2001; Werrett, 1997), phylogenetics (Bowcock et al., 1994), population genetics (Rosenberg et al., 2002) and performing both genome-wide and more local searches for disease susceptibility genes (Ophoff et al., 2002). The repeat array class of microsatellites will make them more suited to different analyses. For example, dinucleotide repeat loci mutate more readily than many tri- or tetranucleotides (Chakraborty et al., 1997), but not all (Nikitina and Nazarenko, 2004), and are highly abundant in the genome. However, tri- and tetranucleotides give less PCR 'stutter' and are often preferred for this reason. 'Stutter' refers to a different length of allele being generated from the template allele during amplification, possibly due to polymerase slippage. Although stable, five and six basepair repeats are generally a lot less frequent and are not as popular for this reason.

SNPs have two major benefits for use in genetic studies: (i) they are found in relative abundance throughout the genome, meaning that whatever region of the genome is being studied a SNP marker will likely reside nearby, and (ii) genotyping SNPs is relatively straightforward making them very amenable to automation and cost-efficiency. As they are less polymorphic and have smaller mutation rates than microsatellites they are more suited to describing variation that has existed in populations for longer periods. To date they have helped explain phylogenetic relationships between human groups (Underhill et al., 2001), illuminate demographic histories of human populations (Reich et al., 2001; Zietkiewicz et al., 2003) and are commonly used in genetic association studies that aim to find genes important in human disorders (Botstein and Risch, 2003). More recently, genome-wide based association studies are being developed (Wang et al., 2005), as are techniques that detect evidence of natural selection across the entire genome (Akey et al., 2002); both of these depend on the abundance, distribution and automated genotyping of SNPs.

1.1.4 Variation in Genetic Expression

One form of non-coding genetic variation subject to an increase in research in recent years is that which is responsible for variable levels of genetic expression. In 1975 the potential role of variation in gene expression levels in the evolution of species was realised when King and Wilson observed that genetically similar species often show a disproportionate level of anatomical differentiation, and that variation in levels of expression may therefore be responsible (King and Wilson, 1975). This early idea has since been reinforced by the discovery (made possible by the completed genome sequence) that the human genome contains fewer genes than previously thought; suggestive that the wide distribution of phenotypic variation seen among humans is due to factors other than protein-altering mutations.

These observations have instigated research in the last four years concerned with determining the extent of naturally occurring variation in expression levels between individuals, and its biological significance. In 2002 Oleksiak *et al* used a type of microarray analysis sensitive to small differences in expression and showed 18% of 907 genes had significant variation in expression levels within a naturally occurring population of *Fundulus* fish. Variation factors fluctuated around 1.5 and some genes varied by a factor of 2 or more. They also went on to show larger variation between two separate populations, each subject to different environments, than within populations for fifteen genes and concluded that some of these differences may be maintained by natural selection (Oleksiak et al., 2002).

In an effort to assess naturally occurring expression variation in humans, Cheung *et al* (2003) performed microarray analysis in 813 genes in 'immortalised' lymphoblastoid cells of 35 unrelated CEPH individuals. For the forty most variable genes variation factors were 2.4 or greater. Notably these researchers showed, via a family based study, a strong heritable component to the differences in expression (Cheung et al., 2003).

1.1.4.1 Role of *Cis*-acting Variants

The next important stage in the study of expression variation is the identification of specific genetic determinants. Research has recorded regulation effects due to *cis*-acting variants *in vivo* in mice (Cowles et al., 2002), and Yan *et al* found similar results in humans (Yan et al., 2002). Yan *et al* (2002) investigated the

association of cSNPs on variable expression levels in thirteen human genes, using lymphoblastoid cells from 96 CEPH individuals and a method based on fluorescent dye termination of RT-PCR products. Six of the thirteen genes had significant expression variation with variation factors ranging from 1.3 to 4.3, present in 3 to 30% of the population dependent on the gene being analysed. This demonstrates the abundance and significance of naturally occurring *cis*-acting variation in gene regulation in humans. They also demonstrated a Mendelian inheritance pattern in expression patterns for two genes.

Hoogendoorn *et al* (2003) wanted to assess the background frequency of promoter variants in human populations, and determine its role in gene expression (Hoogendoorn et al., 2003). They screened 170 promoters (average size approximately 430 bp) selected randomly from the Eukaryotic Promoter Database for novel SNPs, with an aim to explore the frequency and effect of this polymorphism. Of these promoters, 41 contained polymorphism and were cloned into pGL3 vectors upstream of a luciferase gene and transfected into 3 different human cell lines, as part of a luminescence based reporter gene assay. They found that at least 35% of human promoters were polymorphic, and 10% of human genes have functionally relevant variation in their promoters. Although these levels of variation were considered substantial, these projections are likely to be underestimates as the study design was insensitive to less frequent variants.

Together these studies have shown that there is considerable, naturally occurring variation in gene expression levels, that variation is heritable and that at least part of this variation may be due to *cis*-acting factors. Variants that cause variable levels of gene expression may therefore be targets for natural selection and have had important roles in the evolution of humans.

1.2 Human Evolution

There are several important features of the eukaryotic genome common to all eukaryotes. However, there are many more species-specific features that have arisen during the divergence of new species from ancestral lineages. And, accordingly, it is differences in the genome of the common ancestor of us humans and our closest

living relatives, the chimpanzees, which gave rise to the hominid lineage and eventually modern humans.

1.2.1 Origins of Modern Humans

Humans and the chimpanzees share a most recent common ancestor that walked the earth between 5 and 7 million years ago (Mountain and Cavalli-Sforza, 1994), and incomplete fossil evidence allows a contentious evolutionary path to be drawn to present day man. One of the earliest hominids was the Australopethicine *Ardipethicus ramidus ramidus* from 5 MYA, although a new earlier species has also been reported to have existed in the form of *Sahelanthropus tchadensis* (Brunet et al., 2002). It is believed that the gracile Australopethicines gave rise to the Homo genus, and *habilis* was long considered to be the first member (Leakey et al., 1964). However, a more comprehensive inclusion criterion for Homo now suggests that either *erectus* or *ergaster* was the first Homo species evolving around 1.8-1.9 MYA (Wood and Collard, 1999).

Archaeological evidence shows that over time anatomically modern humans (AMH), *Homo sapiens*, made their first appearance between 130 and 160 KYA in Africa (White et al., 2003), separated from early Homo by an increase in globular shape of the skull and a decrease in the retraction of the face (Lieberman et al., 2002). AMH have been found in Asia and Australia between 40 and 60 KYA, and in Europe around 39 KYA (Zilhao, 1999). The earliest modern humans in the Americas are dated to around 13.5 KYA (Cooke, 1998) and the last lands to be colonised were the more remote islands of Oceania around 3.5 KYA (Kirch, 1999; Merriwether et al., 1999) (see figure 1.2).

Culturally, archaeological evidence (such as tools and artwork) suggests that humans were not modernised until 80 KYA in Africa, and perhaps 50 KYA outside Africa (Burenhult, 1993; Henshilwood et al., 2002). Cultural advances are likely to have coincided with the development of modern languages (Ruhlen, 1994).

1.2.2 Mode of Global Colonisation by AMH

By the time anatomically modern humans had reached the Oceanic islands *Homo sapiens'* range stretched over 70% of the earth (reviewed by Jobling et al.,

2004). Migrations were most likely governed by environmental factors such as drastic temperature fluctuations associated with the glacial events of the last 100 KY. However, the mode of globalisation has been a hotly contended issue reduced to two major theories: the 'Out of Africa' model, and the multiregional model. The 'Out of Africa' model proposes that AMH arose in Africa once and then migrated to other continents, perhaps with a small amount of interbreeding with other Homo species. The multiregional model, on the other hand, suggests that AMH evolved from early Homo species in Africa and other species (for e.g. *erectus*) that had migrated out of Africa. As populations evolved over time a lot of movement and gene flow occurred between all populations; meaning AMH evolved from a widely dispersed meta-population of early humans.

1.2.2.1 Genetic Evidence

Many types of evidence have helped determine which of these models is likely to be correct. Palaeontological records show the oldest fossils come from Africa, whereas more recent fossils come from the rest of the world. This seems to support 'Out of Africa', but multiregionalists would argue that this may only show that populations in Africa were bigger or represent the amount of research time spent in Africa compared to other regions.

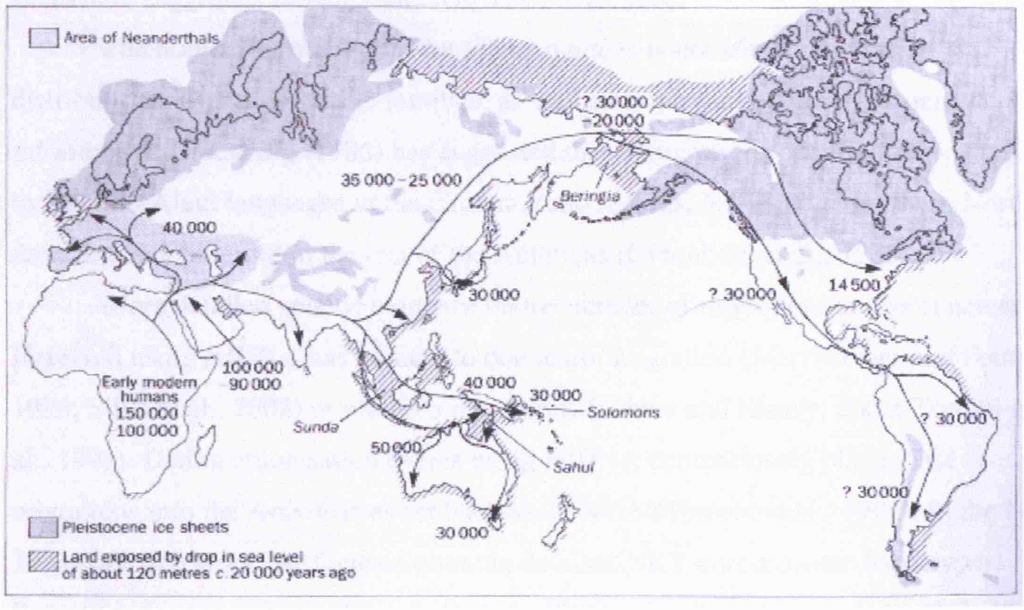
More recently collected evidence helps resolve the issue. For example, molecular data consistently shows African populations to be more genetically diverse than populations from other continents (Bowcock et al., 1994; Jorde et al., 2000; Stoneking et al., 1997; Tishkoff et al., 1998). If modern humans arose in a multiregional fashion then all populations today should be roughly the same age and derived from a population with roughly the same effective population size. This should give no grounds for an increase in genetic diversity of African populations.

Further, dating colonisation attempts based on the distribution of genetic diversity are consistent with archaeological data in showing that modern man was initially present in Africa and then sequentially in Asia, Europe and finally Oceania, and the Americas (Watkins et al., 2001; Zhivotovsky et al., 2003).

These lines of evidence have led to the 'Out of Africa' model being favoured. Further support is provided by linguistic analysis. For example, when the global distribution of human languages is reviewed more language families exist in Africa than in more recently colonised lands and is regarded as support for the 'Out of

Africa' model (Ruhlen, 1994). Researchers now try to resolve the contentious issue of the number of times modern man migrated from Africa (Templeton, 2002; Zietkiewicz et al., 2003).

Figure 1.2. Summary of possible origin and global expansion of anatomically modern humans. (From Jones et al., 1994).



1.2.3 Colonisation of the Americas

Due to the extent of landmass and relatively recent appearance of AMH, the Americas represent an excellent opportunity to find artefacts important in the study of ancient human cultures and technologies. Exactly when and how the Americas were colonised by modern humans has not been indisputably resolved. Discovery of the Clovis spear point in New Mexico (named after the village in which they were found) shows that colonisation had occurred by 12-13 KYA (Cooke, 1998). Evidence also exists for pre-Clovis cultures and the presence of culturally advanced humans in Monte Verde, southern Chile has been suggested at 14.5 KYA and as long ago as 33 thousand years, although this latter result needs confirmation (Dillehay, 1997;

Meltzer, 1995; Meltzer, 1997). It is widely agreed that the main route for the initial settlers was from North-East Asia across the Beringian land bridge; presently submersed by the North Atlantic Ocean and known as the Bering Strait. However, the land bridge has only been accessible twice in times recent enough to be used by modern humans and, in addition, land routes southward from North America have been prevented by an ice cap during the last glacial maximum. The fact that a corridor through the ice cap was not available at the time the Beringian land bridge was accessible suggests a coastal route may have been used.

The number of migrations into the Americas is not clear. Using the distribution of major language families, as well as some limited gene frequency information, Greenberg (1986) has suggested that 3 migrations occurred represented by Eskimo-Aleut languages in the circum-arctic regions, Na-Dene in northern North America and Amerind in the rest of the Americas (Greenberg et al., 1986).

More detailed genetic evidence on the number of migrations is less conclusive. Research using mtDNA has pointed to one major migration (Merriwether and Ferrell, 1996; Silva et al., 2002) or multiple migrations (Schurr and Sherry, 2004; Torroni et al., 1993). Dating colonisation events using mtDNA contentiously places first migrations into the Americas as far back as 29 KYA (Torroni et al., 1994). In the last 10 years much work has focused on more detailed NRY chromosome haplotypes. Results so far support one (Underhill et al., 1996), two (Bortolini et al., 2003; Lell et al., 2002; Ruiz-Linares et al., 1999) or three migrations (Schurr and Sherry, 2004). Less varied are the dating attempts using the Y chromosome which place the arrival of modern humans from around 10 to 14 KYA (Bortolini et al., 2003; Ruiz-Linares et al., 1999), in stronger agreement with archaeological evidence.

It is clear that current research has not been able to definitively explain the colonisation of the Americas. The concept of two major migrations, at least one of which used a coastal route, may best explain the existing evidences. Issues are unlikely to become any easier to resolve genetically with time, however; admixture of American populations and ancestral populations (such as those from Siberia) will likely increase with time.

1.3 Human Diversity

The evolutionary history of our species and demographic events leading to continental colonisations has shaped genetic diversity within our genomes in a population-specific manner. For example, evolutionary factors including genetic drift and different selection mechanisms such as negative and positive selection, background selection and selection sweeps will remove and arrange variation in the genome over time. Populations will be affected by these factors to varying degrees dependent on the environmental stimuli and stochastic demographic events, such as expansions and bottlenecks, to which they have been exposed. Ultimately, this may give rise to population specific genomic structure and genetic variation. How, then, is genetic diversity distributed across modern human populations?

1.3.1 Genetic Diversity

There are several measures of genetic diversity: the proportion of polymorphic loci, heterozygosity and gene diversity (Nei, 1987). The proportion of polymorphic loci considers the number of segregating sites between individuals. Locus heterozygosity refers to the proportion of heterozygotes at a locus, and average heterozygosity refers to the mean of this value over a certain number of loci (see section 4.2.4.1). Gene diversity, on the other hand, is calculated from the expected number of homozygotes based on allele frequencies, which can also be averaged over a number of loci (see section 3.2.4.5). Gene diversity may therefore be better suited to detect diversity in situations where very few heterozygotes are present.

1.3.1.1 Genomic Distribution of Genetic Diversity

Variation most likely to persist in a population will either have little or no effect on gene function. Theoretically, it is straightforward to predict whether coding mutations may affect gene function, by determining whether a synonymous or non-synonymous nucleotide change has occurred, and predicting whether non-synonymous changes are in functionally important regions of the encoded protein. It is less obvious, however, whether non-coding variation will have a significant functional effect. For example phenotypic changes caused by variation in regulatory

regions, achieved by altering gene expression, are likely to be more subtle than coding variation, and will not be acted upon as strongly by selection.

A reasonable assumption might therefore be that the sequence diversity of non-coding regions is greater than that of coding DNA. A genome-wide analysis has shown the average number of polymorphic sites per kb to be 5.9 in sequence 5' to transcription initiation start sites, a larger value than was obtained for coding regions estimated at 3.4 (Stephens et al., 2001). Work done by Hoogerndoorn and colleagues (2003) showed an average nucleotide diversity of 4.9×10^{-4} in promoters (Hoogendoorn et al., 2003). This is actually slightly less than an estimate for entire coding regions, at 5.30×10^{-4} (Cargill et al., 1999). When considering only non-synonymous basepair changes, however, coding diversity drops drastically to 2.56×10^{-4}, and to 0.91×10^{-4} for non-conservative changes (Cargill et al., 1999); substantially lower than for promoter regions.

1.3.1.2 Species Distribution of Genetic Diversity

There are several characteristics of human genetic diversity that are specific to the species. For example, low levels are observed compared to other species with similar numbers and ranges. Average global heterozygosity values have been observed between 0.66 to 0.73 for multi-allelic STRs (Jorde et al., 1997; Jorde et al., 2000; Rosenberg et al., 2002) and lower in bi-allelic systems such as *Alu* insertion polymorphisms (0.25 to 0.26) and RFLPs (0.35 to 0.37) (Jorde et al., 2000; Watkins et al., 2001). This comparatively low genetic diversity in humans, when considering non-human primates, is believed to reflect the relatively young age of the species, as well as a small founding population (Kaessmann et al., 1999; Kaessmann et al., 2001).

Another attribute of human genetic variation is that it is not equally distributed across the globe. Studies consistently find Africa to harbour the majority of genetic diversity, Europe and Asia have intermediate levels, whereas the Americas and Oceania consistently show low levels of global genetic diversity (Bowcock et al., 1994; Cann et al., 1987; Jorde et al., 2000; Rosenberg et al., 2002; Watkins et al., 2001; Zhivotovsky et al., 2003). This trend is consistent with theories that an expansion occurred in Africa before major global emigration occurred or that bottlenecks occurred during or prior to colonisation of non-African regions (Excoffier and Schneider, 1999; Reich et al., 2001; Reich and Goldstein, 1998).

1.3.2 Linkage Disequilibrium

A further indicator of human genetic diversity is genomic linkage disequilibrium. Linkage disequilibrium (LD) is the non-random association of alleles in our genomes and describes the extent to which genomic segments remain intact from the mixing effects of recombination. It can be visualised in figure 1.3.

LD can be most conveniently described if a two locus system is considered, loci A and B, each with two alleles represented as A or a, and B or b respectively. LD is assessed by comparing the observed frequency of haplotype AB (pAB) to that expected if alleles were randomly segregating, determined by the product of allele frequencies, pA and pB. This is known as the linkage disequilibrium coefficient, D:

$$D = pAB - pApB$$

where p is the population frequency (Lewontin, 1964). D can be either positive or negative and is directly dependent on allele frequencies.

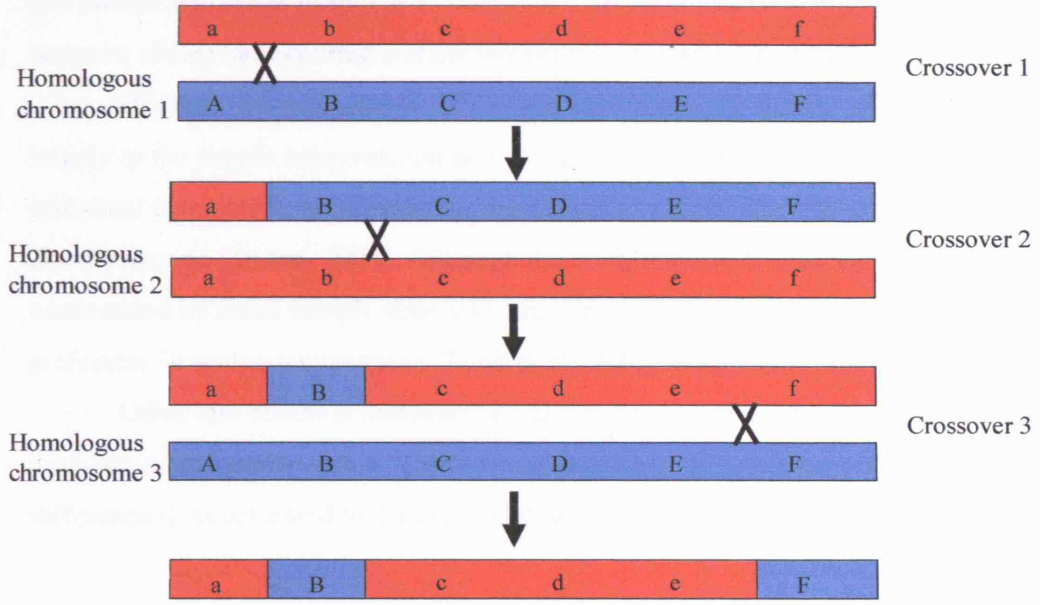
Statistically, linkage disequilibrium exists if the null hypothesis that alleles are segregating independently from one another is rejected, shown if the frequency of haplotype AB is significantly different from the products of the population allele frequencies of A and B. Statistical tests, such as the Chi square or Fisher's exact test (FET) can be used, and if probability values fall below significant values then LD is assumed (Weir, 1996).

D decays by recombination over time and this relationship can be explained by:

$$D_t = (1\text{-}r)^t \times D_0 ,$$

where D_0 is the current value of D, D_t is D in t generations and r is the recombination rate between two alleles. We can see from this equation that LD decreases with time, as a result of increased recombination. As physical distance between two alleles increases so does the chance of recombination, therefore D tends to decrease with increasing distance between loci.

Figure 1.3. Pictorial representation of linkage disequilibrium.



The linked alleles c, d and e have remained intact over three meioses.

1.3.2.1 Measuring LD

In addition to the statistical confirmation of LD, qualitative measures are also meaningful and enable the strength of LD to be assessed and compared across populations. Several measures have been devised, all of which are dependent on D. One of the most popular is Lewontin's D' calculated by:

$$D' = D / Dmax ,$$

where Dmax is the maximum possible value of D when $D \geq 0$, or the minimum possible value of D when $D < 0$ (Lewontin, 1964). D' values range from 0 to 1, where values of 1 reflect complete LD and is achieved if no recombination has occurred between alleles throughout the history of the study sample.

Another useful measure of LD is the statistical association measure r^2, also denoted Δ^2 (Hill and Weir, 1994). It assesses the correlation between alleles and is generated by normalising D by the square root of allele frequencies:

$$r^2 = D^2 / \ pApapBpb$$

and values also range from 0 to 1. Values of 1 are obtained only if no recombination between alleles has occurred and the frequencies of alleles are equal.

Of these two measures of LD, D' has been more popular for gene mapping due largely to the simple interpretation of a D' value of 1. r^2, with more stringent statistical parameters, is favoured for theoretical modelling (Devlin and Risch, 1995; Zondervan and Cardon, 2004). Although the interpretation of both D' and r^2 is confounded by small sample sizes and rare alleles (Weiss and Clark, 2002), r^2 may be preferable in such circumstances (Teare et al., 2002; Wall and Pritchard, 2003b).

Other less common measures of LD that have been used are Levin's population attributable risk δ, Yule's Q and Kaplan and Weir's proportional difference d, as reviewed by Devlin and Risch (1995).

1.3.2.2 LD structure of the human genome

As LD may have important implications in finding genes that have a role in human disease (Zondervan and Cardon, 2004), a lot of work has focused on determining the extent and structure of LD in the human genome (Ardlie et al., 2002). An early estimate based on computer simulations suggested useful LD in the general population is limited to approximately 3kb (Kruglyak, 1999). However, significant LD values have been recorded over much greater distances, up to 100kb, by an analysis of 9 genes over 135kb in a European population (Johnson et al., 2001). It is now understood that levels of LD vary around the genome (Reich et al., 2001). This implies that using average genome-wide estimates may not be appropriate for disease association, which rely on LD at specific genomic regions.

There are several factors believed to influence the patterns of LD in the human genome. These include evolutionary factors such as genetic drift and selection, and demographic factors such as admixture at founding and population dynamics; all of which can generate LD in a population-specific manner. Selection may also generate species-wide patterns of LD. LD may be localised to certain genomic regions, as a result of selection or variable mutation rates; or genome-wide as a consequence of stochastic demographic events. Further, as LD is broken down by recombination, sites of crossovers will have a direct influence on how LD is distributed throughout the genome. It is apparent, therefore, that it is the interaction of many determinants, each working in a specific way, which gives rise to observed population patterns of

genomic LD. Considerable effort has been made within the last ten to fifteen years to understand the causal factors further.

Demographic history

Demographic events in a population's history mould its genetic make-up, including the pattern of LD. Population genetic theory predicts that on average young, small, isolated populations show high degrees of LD, and particularly those that grow slowly. Alternatively, old, large, outbreeding populations should show lower amounts of LD, as should populations that have undergone fast exponential growth (Laan and Paabo, 1997; Reich et al., 2001; Slatkin, 1994).

Admixture

Admixture can be described as the mixing of two populations that were previously isolated from one another to form a hybrid population. Mixing may occur instantaneously or more gradually over time. As well as generating new population allele frequencies, early studies of admixture showed increased levels of allelic association (Chakraborty and Weiss, 1988). The underlying process can be explained as follows: as members of ancestral populations come together their separate genomes represent whole linked genomes; soon this initial large increase in LD breaks down by chromosomal segregation; eventually chromosome specific LD breaks down by recombination. The extent to which admixture influences LD depends on the genetic make-up of the ancestral populations and the dynamics of population mixing. Generally, the greater the divergence of allele frequencies in ancestral populations the stronger the LD generated between these alleles after admixture (Collins-Schramm et al., 2003; Pfaff et al., 2001). Furthermore, the complexity of the population mixing process has also been associated with strength of LD (Pfaff et al., 2001).

Selection

Natural selection may lead to combinations of alleles at different loci. For example it is obvious how an advantageous polygenic trait may cause a non-random segregation of alleles at different genes (Cooke and Hill, 2001). Less obvious is the effect on alleles adjacent to the locus under selection. If selection is strong enough then the target locus may rise to high population frequencies in a short amount of time. As sufficient time has not passed for recombination to break down allelic

associations with the surrounding genomic region, then the selective sweep and genetic hitch-hiking may give rise to high LD (Gilad et al., 2002; Sabeti et al., 2002). Selection can therefore generate local chromosomal regions of LD.

Genome specific patterns

Evidence in the last few years has revealed that recombination may not be evenly distributed across the genome, but rather localised leading to regions with significantly higher rates than the genome average (Daly et al., 2001; Goldstein, 2001; Hey, 2004; Jeffreys et al., 2001; McVean et al., 2004). These regions have been termed recombination hotspots and may span several kilobases. The cause of hotspots is not fully understood, however some evidence suggests that sequence motifs may have a role (Hurles et al., 2004). Hotspots would result in a punctate distribution of LD and effectively partition the genome into discrete blocks (Daly et al., 2001; Jeffreys et al., 2001; Jeffreys et al., 2005). Other research supports a haplotype block structure and existence of recombination hotspots in the genome, but that most blocks are not necessarily separated by hotspots (Wall and Pritchard, 2003a; Wall and Pritchard, 2003b). Togeher these observations have led to the recent development of the 'International HapMap Project', which uses a set of common SNPs to represent the majority of variation in different ethnic groups, exploiting the block-like structure of the human genome (Gibbs et al., 2003). A major aim of this project is to increase the efficiency of finding genes important in complex diseases (Foster, 2004).

1.3.3 Genetic Structure of Human Populations

Many early beliefs, including those of Charles Darwin, were that the human species was divided into many races, with Europeans inevitably coming somewhere near the top of an artificial hierarchy. These views were not based on scientific fact, however, but rather reflected the views of European society at the time. Although research has since shown overall genetic diversity levels to be low in humans, work in the last fifteen years has revealed that the genetic variation that does exist has a global distribution consistent with geographic location.

A useful statistic to describe the degree of population structure is the F_{ST}. The F_{ST} is one of Wright's F statistics concerned with the comparison of genetic variation

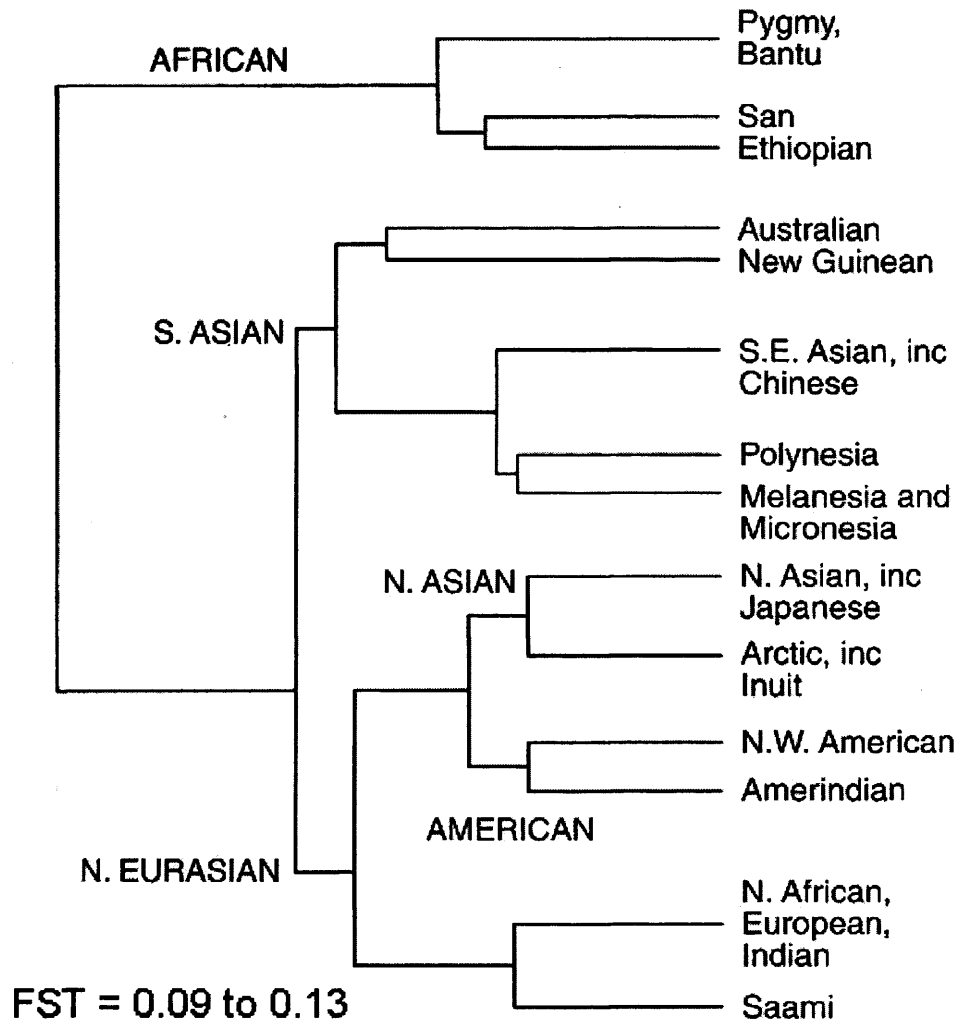
at various population hierarchal levels; the F_{ST} compares the amount of variation that occurs between sub-groups to the variation present within the total sample. There are several ways to calculate F_{ST} and various types of genetic data can be used, however the most straightforward is a comparison of heterozygosities;

$$F_{ST} = (H_t - H_s)/H_t ,$$

where H_t is the heterozygosity of a large meta-population and H_s is the mean heterozygosity between sub-populations. Values vary between 0 and 1; large values indicate that sub-populations from the meta-population are genetically distinct from one another, whereas low values suggest no structure is present.

Estimations of genetic differentiation using F_{ST}s show the level of population structure to be low in humans compared to other species of similar size, with estimates ranging from 9 to 13% (Barbujani et al., 1997; Bowcock et al., 1991; Jorde et al., 2000; Romualdi et al., 2002). This may reflect our recent founding and small amount of time for populations to become genetically distinct. However, differentiation has repeatedly been shown between populations of African, Asian, Oceanic and European origins based on genotype frequencies (Bowcock et al., 1994; Cavalli-Sforza, 1994; Jorde et al., 2000; Rosenberg et al., 2002; Zhivotovsky et al., 2003). Phylogenetic work based on sequence similarity of non-recombining portions of the genome also consistently sub-group the world's population into Sub-Saharan Africa, North Africa, Europe, East Asia, South Asia, the Americas and Oceania (Quintana-Murci et al., 1999; Underhill et al., 2001). It appears, therefore, that the limited amount of variation present in humans can be apportioned into major continental groups (see figure 1.4).

Figure 1.4. Genetic distance tree of human populations based on 120 classical markers. (From Cavalli-Sforza, 1994.)



1.3.4 Distribution of Genetic Variation in the Americas

There are two attributes of the genetic diversity harboured within the pre-Columbian peoples of America that clearly show the importance of evolutionary factors to patterns of genetic diversity. Firstly, diversity levels are consistently shown to be low compared to other parts of the globe (Rosenberg et al., 2002; Zhivotovsky et al., 2003) which may well represent population contractions associated with colonisations. Secondly, some of the highest human population structuring has

occurred within the Americas, with F_{ST} values two to three times higher than for other continents (Mesa et al., 2000; Rosenberg et al., 2002). This is likely due to substantial prevention of gene flow between populations by geographic barriers, such as mountains and ice caps, causing strong tribalisation early on after arrival by initial settlers (Bortolini et al., 2003), reinforced by the development of distinct language groups (Ruhlen, 1994). The establishment of small, isolated populations means that the effects of genetic drift will drive population differentiation more rapidly than for large, interbreeding populations (Mesa et al., 2000).

1.3.5 Implications of the Genetic Structure of Modern Humans

Over the last fourteen years the Human Genome Diversity Project has evolved, the aim of which is a comprehensive analysis of the levels of variation in human genomes and degree of structuring between populations (Cavalli-Sforza, 2005). The global distribution of human genetic variation and human population structure may have practical consequences. For example, established patterns of genetic diversity within a population can be used to help elucidate that population's evolutionary history (Bowcock et al., 1994; Cavalli-Sforza, 1994; Harpending et al., 1998; Jorde et al., 2000; Ruiz-Linares et al., 1999; Tishkoff et al., 1998). Additionally, alleles that influence human phenotypes may differ in frequency between populations thereby influencing the frequencies of phenotypes in genetically distinct groups (Risch et al., 2002), and this argument extends to non-genetic factors that affect human traits. As well as obvious differences in physical appearance between various continental groups, examples include a higher incidence of sickle cell anaemia in countries where malaria is endemic and a higher prevalence of Huntington Disease in Western Europe than in other regions of the world (reviewed by Jobling et al., 2004). Furthermore, Wilson and colleagues demonstrated varying response to disease resistant drugs for groups with different genetic identities (Wilson et al., 2001). Consideration of the unique genetic make-up of populations may therefore be important for the elucidation of genetic and non-genetic components of human traits, and may be particularly
important for common complex, multifactorial traits for which relatively little is known.

1.4 Complex Traits

1.4.1 General

Human traits can be separated into two major types: Mendelian, whose study was pioneered by Gregor Mendel in 1866 and later extended into the 20th century by William Bateson; and non-Mendelian, characterised via the development of biometry by Francis Galton around 1900 (reviewed by Strachan, 2003). It is their underlying mechanisms that distinguish between the two classes. Mendelian traits result from one or a few genes, the mode of inheritance is known and there is a good understanding of trait characteristics such as environmental influence, phenocopy rate and penetrance values of causal genotypes. A non-Mendelian (complex) trait can be any that does not fit into this category, and may be caused by many genes and environmental factors that interact in a specific manner. The genetic background of multi-factorial characters may either be oligogenic, dependent on the combined action of several loci, or polygenic, in which the interaction of many genes is required.

Complex traits can be further classified into continuous and dichotomous forms. Continuous traits, such as height or IQ, are considered polygenic as many genes interact resulting in a spectrum of phenotypic values for a population. Polygenic characters are often described as being quantitative traits and, due to their importance to the farm animal industry, a lot of work has been done by animal breeding centres to determine heritability values and characterise the quantitative trait loci that govern these traits.

Dichotomous traits are those that are either present or absent with no, or little, phenotype distribution. They may be governed by many genetic factors interacting with many environmental factors. Falconer (1981) used a 'causal factor threshold' to describe them, which is explained by the phenotype being expressed only if there is the correct number and array of causal factors; an individual may have all but one factor and still not show the phenotype.

1.4.2 Complex Disorders

One group of complex traits of particular relevance to human health are common human diseases. Like all disease, complex diseases and disorders are traditionally

regarded as being dichotomous. Examples include the cancers which are caused by accumulation of genetic mutations, and diabetes mellitus which, although a genetic component is known for types 1 and 2, may be strongly influenced by environmental factors (see Strachan and Read for reviews). However, unlike their Mendelian counterparts relatively little reliable genetic information has been established for complex disorders. For example, at the Human Gene Mutation Database there are presently (August, 2005) recorded 1863 genes associated with disease (Stenson et al., 2003), the vast majority of which will be for the easier to map, Mendelian diseases. To date it is estimated true genetic associations have been made for only around fifty complex disorders (Lohmueller et al., 2003; Wang et al., 2005); and for multi-factorial disorders these associations will only represent a component of the total genetic picture.

This paucity of genetic knowledge is due largely to the multi-factorial nature of the disorders, as many genes of modest effect may be involved that interact with the environment in complex ways. Gene mapping techniques may not be sensitive enough to detect moderate gene effects, and are further confounded by the heterogeneity of causes: genetic causes for the same disorder may be different across populations, and even within the same population. This large heterogeneity increases the difficulty of finding shared disease loci. What's more, environmental conditions are very likely to be different between populations and families, and possibly between individuals of the same family; very few studies incorporate adequate controls for the environment. Additionally, although complex human disorders are mainly dichotomous in nature, several classes of phenotype may exist with the possibility that different classes result from different genes (Bearden et al., 2004; Reus and Freimer, 1997). As a result, there remains a poor understanding of the underlying mechanisms of complex disorders (Wang et al., 2005; Zondervan and Cardon, 2004).

1.4.2.1 Genetic Models of Complex Disorders

To help in the understanding of their genetic basis, theories have been proposed that explain the allelic architecture of complex human disorders. The allelic architecture refers to the number of susceptibility alleles, the population frequencies of these alleles and the size of their genetic effects (Pritchard and Cox, 2002; Reich and Lander, 2001). Two extreme theories have prevailed: the common disease/common variant (CD/CV) model and the allelic heterogeneity model.

The CD/CV theory suggests that complex human diseases are characterised by several loci each with only a few, relatively frequent, disease-causing alleles that act inter-dependently (Lander, 1996; Reich and Lander, 2001). Such a situation could evolve if negative selection against common disease variants has been relatively weak, so that a disease allele arising early on in a population's history, when the Ne (effective population size) was small, has evaded elimination. If the new mutation has also evaded removal by drift, then relatively high population frequencies (≥ 0.1) could have been achieved (Reich and Lander, 2001; Zondervan and Cardon, 2004). As modern day human populations derive from an expansion in founding populations around 18,000 to 150,000 years (700 to 6,000 generations) ago, insufficient time has elapsed for the original disease alleles to be replaced by novel disease causing variants, which could significantly reduce their frequencies (Reich and Lander, 2001). Examples of diseases caused by relatively common variants include Alzheimer's disease (Corder et al., 1993; Scacchi et al., 1999), bladder cancer (Engel et al., 2002) and type 2 diabetes (Altshuler et al., 2000).

Alternatively, the allelic heterogeneity model (or many rare variants model) states that numerous rare variants, that act additively and independently, confer common disease (Pritchard, 2001; Smith and Lusis, 2002). Purifying selection prevents high population frequencies (above around 0.01) being reached and, consequently, diseases may have large allelic diversities as any new mutation (also likely to be disease-causing) will have a substantial impact on overall allelic architecture (Zondervan and Cardon, 2004). Examples of diseases that fit this model include deep vein thrombosis (Bertina et al., 1994) and Crohn disease (Hugot, 2002).

Which one of these two extreme models best explains the genetic make-up of complex diseases is still a contentious issue. There are strong arguments for both and empirical evidence shows there to be diversity in the allelic architecture across the relatively few complex disorders to have been characterised to date (Pritchard, 2001). It may be that the allelic structure of many diseases may fall somewhere in between these two extremes dependent on the effects of factors such as purifying and positive selection which generate more rare and common variants respectively (Wang et al., 2005); it is likely, therefore, that disorders will require individual characterisation.

As well as frequency, the extent of phenotypic effect (risk factor) of each susceptibility allele is also a very important factor in the genetic make-up of complex disorders. Without knowledge of the number of susceptibility alleles, sizes of gene

effects are very hard to predict but are likely to be small. For example, research has shown that most quantitative traits are likely to be governed by some loci with large phenotypic effects and many with smaller effects (Barton and Keightley, 2002; Blangero, 2004). Furthermore, alleles that have been reliably associated with a disease consistently show relatively low odds ratios, in the range of 1.1 to 1.5 (Lohmueller et al., 2003; Risch and Merikangas, 1996; Wang et al., 2005); where an odds ratio is the ratio of the odds of exposure to non-exposure in cases compared to odds of exposure to non-exposure in the controls (Kirkwood, 1988). The evidence suggests, therefore, that many of the variants that give rise to complex disorders will be of modest effect.

1.4.3 Mapping Complex Disorder Genes

Mapping genes for complex disorders is notoriously difficult. A first step is to confirm a genetic component by, for example, estimating heritability or comparing prevalence of the phenotype in individuals with varying degrees of genetic relatedness to an affected individual. For example the prevalence of schizophrenia in monozygotic twins has been shown to be approximately 50%, 10-15% for dizygotic twins and siblings, and around 0.15-1% in the general population (Jablensky, 2000; Sawa and Snyder, 2002). These values result in a relatively high heritability estimate for schizophrenia of around 80% (Cardno et al., 1999), and together imply a significant genetic contribution to schizophrenia. Once a genetic component has been established there are two alternative methods for identifying individual loci: linkage analysis and association analysis. Both have advantages dependent on the characteristics of the trait being studied.

1.4.3.1 Linkage Analysis

One of the biggest successes of human genetics in the last 20 years has been the use of genetic linkage analysis to determine the genetic basis for a vast array of human diseases (Botstein and Risch, 2003). The theory behind linkage analysis is to determine if a chromosomal segment, bordered by recombination events, has been co-transmitted with a phenotype to members of a pedigree more often than would be expected by chance. The causative gene for a phenotype should lie in the chromosomal region with statistically significant co-segregation. Microsatellites have

been the markers of choice for this analysis due to their high mutation rates, abundance and genome-wide dispersal.

Most successes of this technique have been for Mendelian disorders for which modes of inheritance, disease gene frequencies and genotype penetrances are known, allowing for linkage analysis to be carried out under a set of robust parameters. Examples of disorders for which genes have been cloned using this method include cystic fibrosis (Riordan et al., 1989), haemocromatosis (Feder et al., 1996), nail patella syndrome (Dreyer et al., 1998) and lactose intolerance (Enattah et al., 2002).

If the parameters of a disease are not known modifications can be made to linkage analyses. These include estimating the parameters of the complex disease and applying parametric linkage analysis which relies heavily on the accurate estimation of unknown disease attributes such as penetrances and gene frequencies. Alternatively, non-parametric analysis can be applied which, in theory, removes assumptions about the genetic model of the disease. However, this approach has a marked reduction in power than parametric linkage and independence from the disease model has been questioned (Sham, 1997).

Linkage works best for disorders with a known mode of inheritance, caused by single genetic variants with strong effect. This approach is less well suited for alleles with only modest effects on disease (Cardon and Bell, 2001; Risch and Merikangas, 1996). As such linkage analysis has been less successful for complex disorders which result from several susceptibility loci each contributing only a fraction of the disease-predisposing genetic variance. Successes for complex disorders include type 1 diabetes (Davies et al., 1994), breast cancer (Hall et al., 1990), hirschsprung disease (Lyonnet et al., 1993) and Parkinson's disease (Scott et al., 2001).

Other drawbacks to linkage analysis are the genomic resolution of this technique isn't particularly fine, often initially limited to between 5 and 10cM, or approximately 5 and 10 megabases (Botstein and Risch, 2003). Further, obtaining samples from all individuals in large pedigrees may be inconvenient and add to both the time and cost of a study, and many pedigrees may be required (Hirschhorn and Daly, 2005; Risch and Merikangas, 1996).

1.4.3.2 Association Analysis

It is believed a more robust way of discovering genes with modest effects is to use association analyses that compare allele frequencies between groups of cases and

controls. The 'case-control' test is a classic example where allele frequencies in a sample of unrelated affecteds are compared to allele frequencies in a group of unrelated healthy individuals; statistical differences between groups implies association of the genetic variant with the disease. The genetic variants tested in association analyses may be functional and reside within a candidate gene, and therefore a causal variant will have been identified by a positive result. However, more often it is the association of a marker allele, in linkage disequilibrium with a disease allele, which is tested for in an 'indirect' association analysis.

The benefits of association-based techniques are their ability to detect gene effects of relatively small effect, and the readiness with which a study sample can be collected, compared to linkage analysis in which extensive pedigrees may be required. However, there are drawbacks such as some knowledge of the underlying biological process of a disease is required before a candidate susceptibility locus can be selected, which may be particularly difficult for complex disorders for which relatively little is known. They also rely to some extent on the sharing of relatively common disease alleles, and may be less effective for detecting rare variants (Hirschhorn and Daly, 2005; Wang et al., 2005).

Historically, a major problem with these tests in general is the high rate of false positives that can be attained; i.e., studies in which statistically significant results are achieved due to reasons other than a true association between a genetic variant and disease (Hirschhorn, 2005). For example, in case-control studies this may occur as a result of population stratification; if the study sample used consists of groups of mixed ethnic backgrounds a significant test statistic may only show different sub-population allele frequencies and not a disease association. Multiple hypothesis testing is also a contributor to false positives, especially for studies that include many variants, and refers to cases where many hypotheses are being tested within the same study. In this case probability values need to be lower than conventional values (e.g. 5%) to correct for the probability that more extreme values are expected by chance than in a situation where only one hypothesis is being tested (Risch and Merikangas, 1996). A further source of false positives lies in systematic genotyping errors generating a bias in data collection (Hirschhorn and Daly, 2005).

Several steps can be made to limit the detrimental effects of these confounding variables of association studies (Zondervan and Cardon, 2004). Inclusion of a potential candidate gene can be based on results from other studies, such as linkage

studies (McCauley et al., 2004; Stefansson et al., 2002), thereby limiting the otherwise broad range of possible loci. Considerations can be made to limit the chances of false positives in studies: careful selection of cases and controls from the same ethnic background should alleviate stratification; significance levels of statistical tests can be adjusted to correct for multiple hypothesis testing by, for example, the Bonferroni correction (Weir, 1996); and genotyping can be repeated for confirmation and reliability.

1.4.3.3 Family-based association tests for mapping complex trait genes

Association based approaches for mapping complex trait genes exploit the genetic constitution of a population to associate alleles with a prevalent phenotype. As described case-control studies were the first type of population association study, based on determining if there is a significant difference in the frequency of marker loci alleles between cases and controls. However, the case-control test is prone to error due to population stratification. Since the early 1980s family-based association tests have been developed that control for such errors. These tests rely on small family units within populations and use non-affected family members, or non-transmitted chromosomes, as controls thereby alleviating errors due to stratification. Their major advantage therefore is to retain exploitation of a population's genetic-make-up while simultaneously incorporating statistical controls. Additionally, for rare alleles (with frequencies less than 0.01) only family-based association studies are believed to provide realistic sample sizes to detect an association (Zondervan and Cardon, 2004). Two common and related family-based association tests are the haplotype-based haplotype relative risk and the transmission disequilibrium test.

HRR and HHRR

The haplotype relative risk (HRR) method of Rubinstein et al (1981) and Falk and Rubinstein (1987) uses information from genetic transmissions in trios, made up from one affected offspring and two unaffected parents (Falk and Rubinstein, 1987; Rubinstein et al., 1981). The genotypes in offspring are compared to artificial genotypes consisting of the two remaining parental alleles. These two genotype groups are considered independent samples and tested as unmatched case-control samples. The HRR was modified by Terwilliger and Ott to use allelic transmissions so that for each trio there are two transmitted and two non-transmitted alleles

(Terwilliger and Ott, 1992). These two groups are then considered as independent case-control samples. As this test looks at marker allele and disease allele together it was labelled the haplotype based HRR (HHRR). The null hypothesis tested is d(1-R) = 0, where d is the LD coefficient and R is the recombination fraction, making this method a good test of LD when θ is 0 (Sham, 1997).

The TDT

One of the potential problems of the HRR and HHRR tests is that they use data from one parent in two categories leading to paired observations and ambiguity of results when parents are homozygous. The McNemar test is also interested in transmitted and non-transmitted alleles using trios, but only considers heterozygous parents. This test determines if there is a preferential transmission of one allele over another to affected offspring. It was renamed by Spielman in 1993 as the transmission disequilibrium test (TDT) who described it as a test that 'considers parents who are heterozygous for an allele associated with disease and evaluates the frequency with which that allele or its alternate is transmitted to affected offspring'(Spielman et al., 1993). Because a significant deviation from a probability of 0.5 for each class means marker allele and disease allele are linked, the TDT is a test for linkage as well as linkage disequilibrium. Exactly how results should be interpreted has been debated but, in general, if trios are unrelated then the TDT is a test of LD and linkage. If trios are derived from a single pedigree, in which all cases have the disease allele identical by descent from a common founder, then the TDT tests for linkage only (Sham, 1997). Actual data may represent something between these two scenarios. Segregation distortion may exist within families and confound interpretation of results; however, the TDT can be modified to correct this by considering alleles transmitted to both affected and unaffected offspring.

Multi-marker Haplotypes

A more powerful approach for determining inherited genetic variance associated with a trait or disorder, than looking for individual markers, is to incorporate haplotype data into tests of association. Both the HHRR and TDT can be extended to test multi marker systems for over-transmission of multi-marker haplotypes (Sham, 1997). However, while an increase in genetic information is

incorporated into each test, an unwanted side-effect is a rise in degrees of freedom (Clayton and Jones, 1999; Lazzeroni and Lange, 1998; Zhao et al., 2000).

1.4.4 Improving Genetic Association Studies

A well documented, and unwanted, attribute of association studies is a lack of replication of positive results between studies. This may well be due to critical features of the genetic basis of the disorder which remain unknown for many common diseases, such as frequency and effect size of susceptibility alleles, and the degree of genetic heterogeneity. These factors will have a direct influence on the power of studies to replicate findings, regarded as a necessary prerequisite before positive results can be considered reliable. In the light of a poor genetic picture, several considerations can be made to enhance the repeatability of studies.

1.4.4.1 Implications of Complex Disorder Allelic Spectra

Knowledge of the underlying allelic spectra of common diseases will have important implications for gene association attempts. The frequency and size of genetic effect of susceptibility alleles influences the efficiency of tests at detecting a true positive signal. For example, association based methods will be more effective if relatively common alleles are shared amongst a population, relying somewhat on the CD/CV hypothesis (Pritchard, 2001; Reich and Lander, 2001; Risch and Merikangas, 1996). The genetic architecture will have a direct effect on the sample sizes required and, consequently, the power of tests, as a positive correlation exists between power and sample size required for a given genetic effect (Wang et al., 2005; Zondervan and Cardon, 2004). Additionally, for indirect tests of association which rely on LD between a marker allele and the disease allele, the frequency of the marker allele and the strength of LD are also considerations; in which case 100% LD between marker and disease alleles at equal frequencies is ideal (Zondervan and Cardon, 2004).

1.4.4.2 Phenotype Class

In the absence of knowledge of the genetic make-up of complex disorders several considerations can be taken to enhance chances of finding causal variants. A major confounding problem lies in their multifactorial nature, as they arise from a combination of genetic and environmental factors. This may give rise to a small

distribution of phenotype classes, caused by different combinations of risk factors. If different classes are included in a study, positive signals could be diluted leading to type II errors; i.e., not detecting a true effect. It is important, therefore, to accurately define disease phenotypes to ensure that all cases in a study share a precise phenotype class (Reus and Freimer, 1997).

1.4.4.3 Genome-wide and gene-wide association studies

In 1996 Risch and Merikangas postulated that, with the advent of complete genomic sequence, a catalogue could be made of common SNPs present in the entire human genome. This, along with efficient genotyping techniques, would allow genome-wide association studies to be carried out, preferable to gene-based approaches as no prior knowledge of underlying mechanisms need be assumed and the majority of all genetic variation present can be tested within the same study (Lander, 1996; Risch and Merikangas, 1996).

The concept of genome-wide association was first considered using SNPs located within genes. However, there is also considerable support for the idea of genotyping SNPs that represent the entire genomic sequence and not just the genes, with the obvious advantage of screening more of the genome for potentially important sequence variation. This approach is dependent on linkage disequilibrium and a haplotype block structure of the human genome, with relatively common haplotypes being shared between individuals (Daly et al., 2001; Gabriel et al., 2002). Although substantially more SNPs are required for coverage of the entire genome, some key SNPs within blocks ('tag' SNPs) may represent the majority of variation present and significantly reduce the number of genotypes required (Daly et al., 2001; Johnson et al., 2001). However, as this method is designed to test for common variation, important rare variants may be overlooked. By focussing on SNPs located only in genes more resources can be dedicated to SNPs that are likely to be functional.

More recently a 'genewide' approach has been proposed, which may be preferable for replication of positive association studies (Neale and Sham, 2004). A genewide approach aims to account for all variation in a candidate gene, in local populations, to demonstrate association of the entire gene as opposed to a marker allele or haplotype. By doing so the problems that occur in replicating results between populations as a result of unknown parameters, such as the allelic spectra of disorders, may be removed (Neale and Sham, 2004).

1.4.4.4 Promoter Polymorphisms as Candidate Loci for Complex Traits

The importance of promoter variation as a target for natural selection in population differentiation and evolution has been highlighted by research that has shown: a high level of naturally occurring variation in genetic expression (Cheung et al., 2003); evidence of the inheritability of variable expression levels (Yan et al., 2002); the direct link between promoter allele and expression phenotype (Hoogendoorn et al., 2003; Yan et al., 2002); and the relatively high nucleotide diversity of promoter sequence (Hoogendoorn et al., 2003). Promoter variants may reach relatively high population frequencies, as phenotypic effects are likely to be modest compared to coding variation, which may prevent its production altogether. There is good evidence to support this (Hoogendoorn et al., 2003; Yan et al., 2002), however it is likely many rare regulatory variants also exist (Hoogendoorn et al., 2003). Together, this data suggests promoter variation may be especially important in complex traits and disorders, in which phenotypes may arise due to common alleles, or the culmination and interactions of several genes, each of modest effect.

Despite this, there have been limited attempts to associate functional promoter variants with common complex disorders. Work that has been done includes a review by Mackay *et al* (2002) who observed that, in Drosophila, QTLs are often associated with polymorphisms in non-coding regions (Mackay, 2001). When Bray and colleagues assessed the expression activity of a COMT haplotype consisting of three SNPs which had previously been implicated in schizophrenia, they found that expression was significantly reduced compared to more common haplotypes (Bray et al., 2003). While one SNP was coding, the other two were intronic and in the proximity of the 3' UTR, respectively. This demonstrated that non-coding alleles associated with disease may detrimentally affect gene function through a *cis*-acting effect on expression. However, the SNPs they identified may not directly affect expression, as their study did not distinguish between the functionality of these SNPs and linkage disequilibrium with other causative variants. Following on from this (and a previous study of theirs), Buckland *et al* (2004) analysed promoter activity of genes differentially expressed in the brains of schizophrenic patients. Of twenty-eight polymorphic promoters, they demonstrated significant functional differences between the haplotypes of eight (Buckland et al., 2004; Hoogendoorn et al., 2003). These results are encouraging and provide good justification for including functional promoter variation in future studies of complex traits and disorders.

1.4.4.5 Useful populations for gene mapping

Heterogeneity of common, complex diseases is a strong likelihood; the same phenotype may be caused by different combinations of risk factors between individuals. For this reason genetically homogeneous populations, such as isolates, that reduce genetic heterogeneity of a disorder, and therefore give stronger signals, are preferential for association studies (Peltonen, 2000; Wright et al., 1999). Isolates tend to be genetically and environmentally homogenous, with a higher probability that allelic variants are identical by descent. This will increase the chance of finding genetic factors due to a high frequency of population-specific susceptibility alleles. As most of these alleles will have a common founder the amount of association masking by identity by state alleles will be limited (Varilo and Peltonen, 2004).

Isolates may also be advantageous due to the nature of LD within them. As variants for complex disorders are only of modest effect they may persist in populations for long periods. Plenty of time will have passed for recombination to breakdown LD surrounding susceptibility alleles. This means for indirect association analyses, isolated and young populations with large amounts of LD might be more suitable for the initial stages of gene discovery (Laan and Paabo, 1997; Wright et al., 1999).

Examples of population isolates that have been used to successfully locate disease genes include populations from Finland, Sardinia, the French Canadians and Iceland. In Iceland the deCODE company was developed to study the genetic make-up of the Icelandic population, in order to find genes associated with complex traits and disorders. Comprehensive genealogies have been established dating back to the settlement of the country. Using this knowledge researchers at deCODE have been able to associate genes with twenty-eight common complex traits such as prostate cancer, obesity and longevity; and isolate genes for twelve disorders including asthma and schizophrenia (www.decode.com).

As our understanding of the underlying genetic mechanisms of complex disorders and traits increases and genotyping methodologies become ever more efficient, the power and effectiveness of association studies will likely increase. The ideal situation would be to undertake comprehensive genome-wide re-sequencing to directly identify sequence variants. However, sequencing technologies have not yet developed substantially to make this a feasible option for the majority of researchers,

either economically or heuristically (Shendure et al., 2004). Until such time, carefully designed association studies may remain the most effective means to identify the genetic basis of common disorders and complex traits.

1.5 Aims

There exists a substantial level of genetic variation in humans. As yet the functional consequences of this variation are not fully understood. My research has focused on the characterisation of human genetic variation and assessing its implications and functional consequences. I have approached this in three ways;

(i) A lot is known about variation that alters coding sequence. However, less is known about non-coding variation, which could have significant functional repercussions. Here, I aim to characterise polymorphism in the promoters of a select group of genes involved in serotonin transmission, which may have an important role in psychiatric traits and disorders.

(ii) The genetic make-up of populations is shaped by demographic and evolutionary history. An example of a geographic population isolate with a unique demographic history is Antioquia, from North-West Colombia. I aim to assess the autosomal genetic make-up of Antioquia using a large collection of SNPs located in seventeen genes. Gene diversity, LD patterns and population structure will be assessed in Antioquia and four carefully selected parental populations. As well as further define the evolutionary history of Antioquia, this information will help determine how best to use this population in genetic association studies of common human diseases.

(iii) Identifying genes with a role in complex traits is difficult, and psychiatric disorders are particularly challenging. One example is bipolar affective disorder, for which previous research both supports and fails to show a role of the serotonin transporter (*SLC6A4*). I aim to investigate the role of *SLC6A4* in BPAD in two Latin-American population isolates. Variation across this gene will be comprehensively assessed in Antioquia and a genetically similar population from the Central Valley of Costa Rica, and any variation common to bipolar patients will be identified.

CHAPTER 2: SEQUENCE VARIATION AT PROMOTER REGIONS IN GENES OF THE SEROTONIN (5-HT) PATHWAY AND ITS EFFECT ON GENE EXPRESSION

CH. 2: SEQUENCE VARIATION AT PROMOTER REGIONS IN GENES OF THE SEROTONIN (5-HT) PATHWAY AND ITS EFFECT ON GENE EXPRESSION

2.1 Introduction

2.1.1 The Serotonin Neurotransmission Pathway is Implicated in Behavioural Disorders

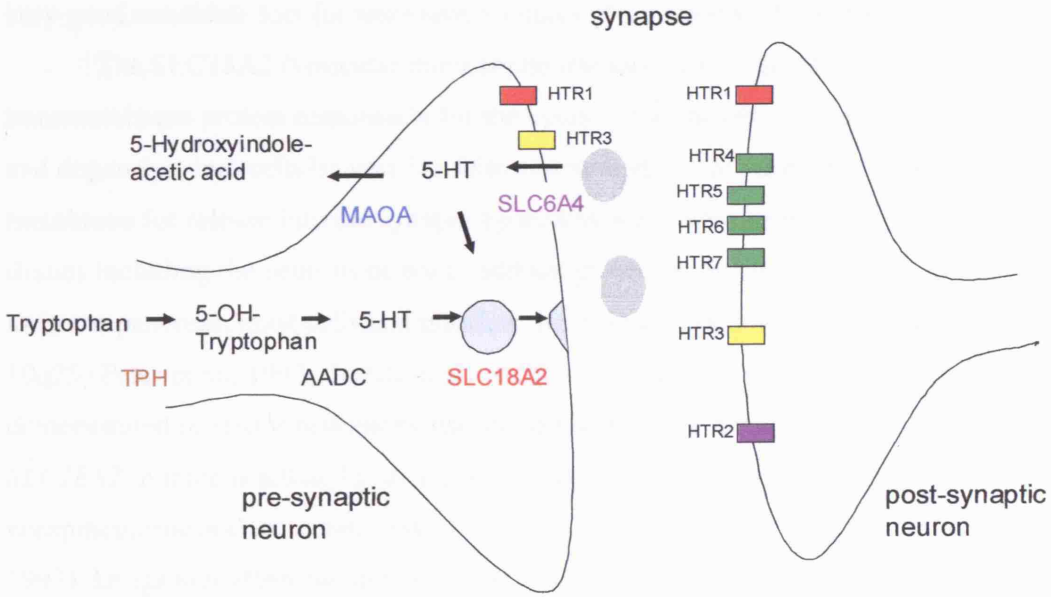
Behaviours are common forms of complex, multifactorial traits influenced by both innate susceptibility loci and proximate environmental stimuli. Behavioural disorders, such as neuropsychiatric diseases, can be particularly debilitating and, since their determinants have remained largely elusive, a lot of effort has focused on their understanding. The serotonin neurotransmitter (5-HT), involved in the inter-neuronal transmission of information, likely plays an important role.

The serotonin neurotransmitter is one of the most widely distributed biogenic, monoamine neurotransmitters throughout the central nervous system. It controls pre- and post-neuronal synaptic gene expression upon activation of receptor molecules which regulate an assembly of effector molecules and proteins (Lesch, 2001a). Serotonin is implicated both in early morphogenesis during embryonic development and, later in developmental stages (even adulthood), in neurogenesis and connectivity of the cerebral cortex (Lesch, 2001b). As such a vital component in the communication of environmental stimuli the system is associated with many functions including sleep, appetite, pain reception, neuroendocrine regulation and mood (Lucki, 1998). It may also control rational thought during consciousness (Donaldson, 1998).

There are many genes in the serotoninergic neurotransmission pathway, which interact to synthesise 5-HT and regulate its activity (figure 2.1.1). Several lines of evidence suggest roles for the transmission pathway genes in neuropsychiatric behaviours: levels of synaptic serotonin have been shown to affect mood, for example a low level of serotonin has long been linked with low mood (Owens and Nemeroff, 1994); they are a target for many psychoactive drugs such as antidepressants, anorectic drugs and recreational drugs (Anderson, 2004; Uhl et al., 2000); altercations of protein activity and distribution have been visualised in individuals affected with psychiatric disorders (Drevets et al., 2000; Neumeister et al., 2004); and observed abnormal behaviour patterns in animals with altered serotinergic function, for

example mice lacking HTR1A receptors show an increase in anxiety (Gross et al., 2002; Parks et al., 1998). Accordingly, the serotonergic genes, responsible for the normal regulation of synaptic serotonin, are popular candidate genes for researchers of psychiatric disorders.

Figure 2.1.1. The serotonin neurotransmission pathway.



Serotonin (5-hydroxytryptamine, 5-HT) is generated in the pre-synaptic neuron from the tryptophan amino acid precursor, under the action of trytophan hydroxylase and aromatic l-amino acid decarboxylase enzymes. Upon synthesis, 5-HT is transported to the synaptic membrane by SLC18A2 and released into the synapse where it stimulates post-synaptic receptor molecules. Activity of 5-HT is terminated by re-uptake by SLC6A4 into the pre-synaptic neuron in a negative feed-back mechanism. It is then either recycled or catabolised by MAOA. 5-HT: serotonin; TPH: tryptophan hydroxylase; AADC: aromatic l-amino acid decarboxylase; SLC18A2: vesicular monoamine transporter; SLC6A4: serotonin transporter; MAOA: monoamine oxidase A; HTR: 5-hydroxytryptamine receptor.

2.1.2 Candidate Genes from the Serotonin Neurotransmission Pathway

Based on function, some serotoninergic genes may represent better candidate genes than others. For example, TPH (tryptophan hydroxylase) enzyme converts tryptophan to 5-hydroxytryptophan in the neuronal synthesis of serotonin, and is

known to be the rate-limiting enzyme in this process (Frazer and Hensler, 1999). A lot of work has been done to try and associate one isoform of the gene, *TPH1*, with behavioural disorders, resulting in little reproducible success. Another form of this gene, *TPH2*, has recently been discovered, mapped to chromosome 12q15 and found to be expressed in the brain in place of *TPH1* (Walther et al., 2003). This suggests that *TPH2*, and not *TPH1*, generates the serotonin precursor, 5-hydroxytryptophan, in the central nervous system. Functional polymorphisms in *TPH2* would therefore make very good candidate loci for association studies of psychiatric disorders.

The SLC18A2 (vesicular monoamine transporter member 2) protein is a transmembrane protein responsible for the uptake of monoamines, including serotonin and dopamine, into cellular vesicles followed by transportation to the synaptic cell membrane for release into the synapse by exocytosis. Expression occurs in many tissues including the neurons of brain, adrenal gland, intestines and stomach, as well as in the pancreas, mast cells and platelets. Its gene has been mapped to chromosome 10q25 (Peter et al., 1993; Surratt et al., 1993). Transgenic experiments have demonstrated *in vivo* functionality and importance: while homozygous knock-out of *SLC18A2* in mice is lethal, heterozygous knock-out approximately halves dopamine, norepinephrine and serotonin content in the brain (Takahashi et al., 1997; Wang et al., 1997). Drugs that affect the nervous system such as amphetamines and cocaine are believed to act on this transporter, and activity of *SLC18A2* has been shown to be drastically reduced by cocaine abuse (Little et al., 2003; Uhl et al., 2000). This gene has also been implicated in cardiovascular dysfunction (Uhl et al., 2000). Together, this data highlights the biological significance of functional regulatory variation in SLC18A2.

One gene to receive much attention has been the serotonin transporter (*SLC6A4*) due to its direct effect on levels of synaptic serotonin. The serotonin transporter mediates the re-uptake of 5HT via Na+ dependent ion exchange from the synapse. It is believed to be the site of action of SSRI (selective serotonin re-uptake inhibitor) antidepressants which inhibit the reuptake of serotonin thereby maintaining synaptic serotonin levels. The gene maps to chromosome 17 q11.1-q12, has fourteen exons and spans approximately 40kb (Ramamoorthy et al., 1993). Characterisation of *SLC6A4* sequence has revealed two functional polymorphisms, one of which lies adjacent to the core promoter (the LPR, length polymorphic region), both of which are known to alter expression levels and are described in more detail in section 4.1. There

have been numerous attempts to associate *SLC6A4* with neuropsychiatric disorders; results remain largely inconclusive (Lesch et al., 1996; Schinka et al., 2004).

The HTR1A receptor is a G-coupled transmembrane receptor that is found both presynaptically and postsynaptically. The presynaptic receptors, acting as autoreceptors, suppress the firing rate of the serotonergic neurons and restrict the release of 5HT into the synapse via an inhibitory feedback system; whereas postsynaptic receptors inhibit serotonin induced gene activation, as well as reduce postsynaptic neuron firing rates (Barnes and Sharp, 1999; Lesch, 2001a). Density of these receptors is greatest in limbic regions of the brain and they have been associated with neurological development and behaviour (reviewed by Barnes and Sharp, 1999). The *HTR1A* gene consists of one 1269bp exon, mapping to chromosome 5q11.2-q13 (Kobilka et al., 1987). Interestingly, modifying this gene can cause behavioural disorders in other mammals (Heisler et al., 1998; Parks et al., 1998; Ramboz et al., 1998). In particular, a study by Gross *et al* (2002) showed a direct relationship between *HTR1A* and anxiety in mice, highlighting expression in the forebrain during early postnatal development (Gross et al., 2002). For these reasons this receptor has been a target gene for many behavioural disorder studies in humans and a role has been both supported (Drevets et al., 2000; Heisler et al., 1998) and rejected (Curtis et al., 1993; Vincent et al., 1999).

A lot of work to date has focused on associating coding variants with psychiatric traits; however, the importance of the level of serotoninergic gene expression has been recognised and suggested as a potential target for novel drug design (Lesch, 2001a). Evidence now exists to support the inclusion of regulatory variation as some genetic associations of psychiatric disorders have implied non-coding regions (Curran et al., 2005; Mynett-Johnson et al., 2000; Shifman et al., 2002). Global expression studies, concerned with assessing expression differences in a wide range of genes between diseased and healthy individuals, have revealed different expression profiles in brain tissue of deceased schizophrenic individuals and unaffected controls (Hakak et al., 2001). Two follow-up studies are of particular interest. In one, a *COMT* haplotype that was strongly associated with schizophrenia and included two non-coding SNPs was later found to significantly reduce genetic expression in the brains of schizophrenics (Bray et al., 2003). In the other, promoter sequence of the differentially expressed genes in schizophrenics, as identified by

Hakak et al (2001), was screened for variation and the functionality of the variants discovered was tested and confirmed (Buckland et al., 2004). However, relatively little work to date has attempted to identify variants in regulatory DNA that have a direct effect on gene expression, and then proceed to associate these variants with a trait or disorder.

The purpose of this study is to screen *TPH2*, *SLCA8A2*, *SLC6A4*, and *HTR1A* for novel SNPs in their core promoters and to assess whether the new found variation is functional using an *in vitro* transcription assay. Although removed from a natural cellular environment, *in vitro* assays allow *cis*-acting regulatory variation to be directly tested for functionality. Such methodology has been used successfully in the past to characterise important promoter based polymorphism (Heils et al., 1996; Sabol et al., 1998). This work could reveal causative candidate alleles that may help in the understanding of neuropsychiatric behaviours and disorders.

2.2 Methods

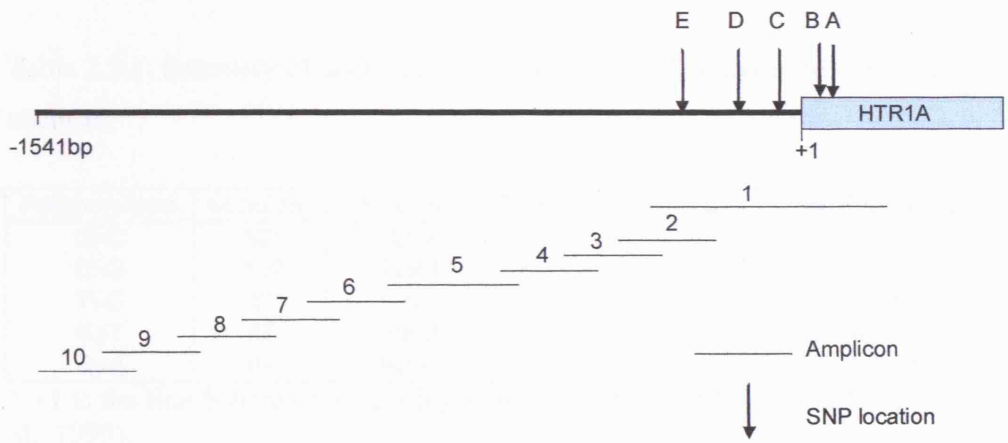
2.2.1 Genes Selected

Four genes from the serotonin neurotransmission pathway were selected for research. These are the presynaptic *TPH2*, *SLC18A2* and *SLC6A4* genes. The remaining gene is the *HTR1A* serotonin receptor. Additionally, *MAOA* was used as an experimental control in the gene reporter assay.

2.2.2 Screening for polymorphism in HTR1A

HTR1A was screened for polymorphism at UCL, with the aim of performing functional analysis on any variants discovered at the Freimer lab in UCLA. Two methods were employed for variation detection. Initially, SNP databases were mined for variation in the 5' region of *HTR1A* and then genotyped in 32 Antioquian BPI affecteds and 3 healthy unrelated Spanish controls. Novel variation was then searched for in sequence extending approximately 1.5kb upstream to the first coding nucleotide of *HTR1A* using single strand conformation polymorphism (SSCP) in the same sample. This region was selected as it not only covers those areas believed to contain important regulatory recognition motifs (Parks and Shenk, 1996), but it also reaches into less well-characterised sequence. Figure 2.2.1 summarises the regions of *HTR1A* screened for variation.

Figure 2.2.1. HTR1A DNA fragments selected for variation screening.



The number above each fragment corresponds to a primer pair. A,B,C,D, and E are SNPs used in RFLP analysis of documented variation; A: 64 G>A, B: 47 C>T, C: -51 T>C, D: -152 C>G, E: -321 G>C.

RFLP of documented SNPs

RFLP (restriction fragment length polymorphism) assays were employed to genotype the documented SNPs. A 600bp fragment that reputedly contains three coding SNPs and three SNPs in putative regulatory sequence was selected for further analysis, and includes 380bp upstream to the start codon and 220bp of the coding sequence. This was amplified using primers designed from the 'Primer3' program (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3) and sequence obtained from GenBank at NCBI. The forward primer is HTR1A-1F 5'GCTTCTCGGTTCTAGATATTTC, and the reverse primer is HTR1A-1R 5'GATAATTGGCCACGTTCTGC. PCR conditions were optimised using DNA from a Spanish population (Valencia). For a 25µl reaction conditions are: 40-100ng of DNA, 1.5mM Mg^{2+}, 1x PCR buffer, 200µM dNTPs, 10pmol each primer and 1Unit(U) of Taq polymerase (Promega, UK). The amplification program comprised of a predenaturation step of 3min at 94°C, followed by 30 cycles each consisting of 30sec/94°C, 30sec/55°C, 30sec/72°C and a final extension step of 10min at 72°C, using an MJ Research PTC-200 Peltier Thermal Cycler. 7µl PCR product was digested using 2 Units of enzyme (New England Biolabs), 1x buffer, 0.1µg/µl BSA and H_2O in a 20µl reaction for a minimum of 4 hours. Electrophoresis was carried out

on digested products using 100 volts for 90 minutes on a 2.5 % agarose gel. SNP data is summarised in table 2.2.1.

Table 2.2.1. Summary of documented SNPs and RFLP assays used in HTR1A analysis.

Polymorphism	Location*	Enzyme	Digestion Temp (°C)	Allele Frequency
G>C	-321	NlaIV	37	99.5/0.5[a]
C>G	-152	HaeIII	37	99.5/0.5[a]
T>C	-51	HinfI	37	99.5/0.5[a]
C>T	47	TspRI	65	96.0/4.0[a]
G>A	64	MspA1I	37	99.32/0.68[b]

* +1 is the first 5' base of the coding sequence. [a](Kawanishi et al., 1998), [b](Nakhai et al., 1995).

Discovery of Novel Variation

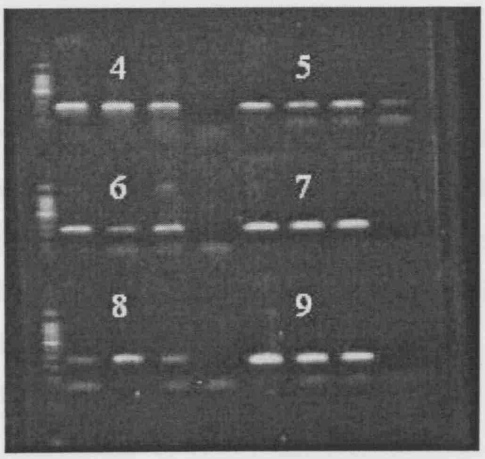
For the detection of novel variation PCR primers were designed for overlapping fragments, approximately 200-250bp in length, for 1329bp of sequence with coordinates -231 to -1560 relative to the first coding basepair, obtained from GenBank. The overlapping primers were designed simultaneously using the ePrimer3 program at the UK Human Genome Mapping Project, Resource Centre (HGMP-RC) (see table 2.2.2). Parameters were set so that melting temperatures fell between 50°C and 60°C and the primer size was between 20 and 22bp. This allowed PCR optimisation to occur under similar conditions and for most primer sets optimisation was reached in a 25μl reaction with: 40ng of DNA, 2mM Mg^{2+}, 1x PCR buffer, 200μM dNTPs, 9pmol each primer and 1Unit(U) of *Taq* polymerase (Promega, UK). For primer pair 8F/8R 80ng of template DNA was required. The amplification program comprised of a predenaturation step of 4 mins. at 94°C, followed by 35cycles each consisting of 30sec/94°C, 30sec/53°C - 55°C, 30sec/72°C and a final extension step of 10 mins. at 72°C (see table 2.2.2), using an MJ Research PTC-200 Peltier Thermal Cycler. PCR primers are shown in table 2.2.2, and the success of the simultaneous primer design for the overlapping fragments for SSCP analysis is shown in figure 2.2.2.

Table 2.2.2. Primers used for SSCP analysis of *HTR1A*.

Primer	Primer Sequence	Fragment Coordinates	Annealing Temperature
HTR1A-2F	5' GGAGAGAGGGAAGGAAGGAA	(-)463 - (-)444	55°C
HTR1A-2R	5' AAACTGGAGTTGGCCTGAAA	(-)250 - (-)231	
HTR1A-3F	5' AAAGGGAAGTGAGGAGGAAGA	(-)562 - (-)542	55°C
HTR1A-3R	5' GGAGTCTCCCCACTAGCAAA	(-)366 - (-)347	
HTR1A-4F	5' AATGCAAAGACGCTGAGCTA	(-)708 - (-)689	55°C
HTR1A-4R	5' CTCTCTCCCCCTCTCTCCTC	(-)475 - (-)456	
HTR1A-5F	5' TGGGATGCTGACACGATTTA	(-)902 - (-)883	55°C
HTR1A-5R	5' GGAGTGCCTCTTTCCTCTGG	(-)740 - (-)721	
HTR1A-6F	5' CGTTGTTCGTTTGTTTTTGG	(-)1109 - (-)1090	53°C
HTR1A-6R	5' TTCCAACTACTCCTTGCCTCA	(-)865 - (-)845	
HTR1A-7F	5' CCTCTCCTTGTCCTTTGACAC	(-)1259 - (-)1239	55°C
HTR1A-7R	5' AAAAAGCTACCTCCGTTCTCG	(-)1046 - (-)1026	
HTR1A-8F	5' AGCCACAAAGCTATGGGAAG	(-)1387 - (-)1368	55°C
HTR1A-8R	5' GTCCAGCCTTACTCCCTCAG	(-)1171 - (-)1152	
HTR1A-9F	5' GGTTTGCAGGCTCTGGTAAG	(-)1483 - (-)1464	55°C
HTR1A-9R	5' CGTGTCAAAGGACAAGGAGAG	(-)1258 - (-)1238	
HTR1A-10F	5' AGGTGCTGAACCCAGTTTCT	(-)1560 - (-)1541	55°C
HTR1A-10R	5' ACTGCCACTTCCCATAGCTT	(-)1380 - (-)1361	

Positions are relative to the first 5' coding base, +1.

Figure 2.2.2. Simultaneous PCR optimisation of 6 different primer sets for SSCP analysis.



1.5% agarose gel run for 1 hour at 100 volts. Numbers refer to primer set. Migration is occurring from top to bottom.

SSCP was performed under the following conditions: 1μl of PCR product and 3μl water were added to 4μl of denaturing solution containing 80% de-ionising formamide, 10mM EDTA (pH 8.0), 1 mg/ml xylene cyanol FF and 1 mg/ml bromophenol blue. Samples were denatured at 98°C for 8 minutes, chilled on ice and then loaded with a 1 in 4 dilution of non-denatured 100bp ladder on a 10% polyacrylamide gel (37.5:1 polyacrylamide:bisacrylamide), 150mm x 140mm x 1mm. Electrophoresis was carried out under 2 different conditions: 300 volts at 21°C for 4 hours, and 400 volts at 4°C for 3 hours. Where necessary, glycerol was added to a final concentration of 2.5 % to enhance separation. Strands were viewed using silver staining and gels subsequently vacuum dried onto 3MM paper at 80°C for 1 hour. If variation was detected, DNA sequencing would have been employed for characterisation; however there was no obvious variation in SSCP patterns.

2.2.3 Screening for Polymorphism in the Pre-synaptic Genes

48 reference DNA samples of Caucasian and 48 of African American origin were screened for novel variation in core promoters by PCR amplification and DNA sequencing. Primers were designed to amplify sequence that had already been shown to contain the core promoter (Boularand et al., 1995; Heils et al., 1995; Xu et al., 1997). For *TPH2*, the core promoter structure was assumed to be similar to *TPH1*; which seems reasonable as they carry out the same biochemical task, and their protein sequence is 70% homologous in humans (Walther and Bader, 2003). PCR was carried out in a 20μl reaction containing 10μl 2 x AmpliTaq Gold MasterMix (ABI, Redwood City, California, USA), 10 pmol of forward and reverse primers, 50 ng template DNA and brought up to 20μl with distilled H_2O. Specific PCR conditions and promoter fragment characteristics are summarised in table 2.2.3. Thermocycling followed a basic protocol of initial denaturation step of 94°C for 5 mins., 30 cycles of 94°C for 30 secs.; 56°C for 30secs.; 72°C for 1 min., followed by a final extension of 72°C for 10 mins. in a Perkin Elmer 9700 GeneAmp® PCR System thermocycler. Success of PCR was checked by electrophoresis. 5μl of product and 2μl of loading buffer were loaded into a 2% agarose gel, made with ethidium bromide, and a current of 90 volts was applied for 45 mins. Products were visualised on an ultra-violet light source (FBTIV-88 Transilluminator, Fisher Scientific).

PCR products were cleaned for sequencing enzymatically by adding 0.2 units exonuclease I and 2 units of shrimp alkaline phosphatase per sample (USB, Biochemicals, Cleveland, Ohio, USA) to the remaining PCR product, and digested for one hour at 37°C, followed by a deactivation step of 15 minutes at 80°C, in the Perkin Elmer thermocyclers. The sequencing reaction was performed using the dideoxy chain termination method (Sanger et al., 1977) in both the forward and reverse directions using PCR primers, and sequences read on an ABI 3700 sequencer using infra-red fluorescence technology. Sequencing was done at the Macrogen Inc. labs in South Korea. Polymorphisms were detected by aligning sequences using the Sequencher program.

Table 2.2.3.Promoter fragments for sub-cloning and specific PCR conditions.

STR locus	Fragment Size	Position (rel to ti)	Primer sequence (5' to 3')	Specific PCR conditions
TPH2	650	-569 to +81	F: ACTGGAAGAGTGGAATTGGAA R: TGGCGGAGATTGAGAGGAAG	
SLC18A2	537	-497 to +40	F: TGCAAAGGGTGGCTTCTTCA R: GCAGTGGGCTCCGTCAGT	8% DMSO
SLC6A4	649	-535 to +114	F: CAGTCAGATAAACGCATGGG R: TTCGAGGCGGAGAGGAAAAG	4% DMSO
MAOA	1252 (low) 1282 (high)	-1236 to +16 -1266 to +16	F: GCTCCAGAAACATGAGCACAAACG R: GGCTGACACGCTCCTGGGTCGTA	8% DMSO

Promoter fragment size, position relative to transcription initiation site and primer sequence used for amplification. ti: transcription initiation.

2.2.4 Reporter Gene Assay

2.2.4.1 Plasmid Constructs

Promoters of *SLC18A2*, *TPH2*, *SLC6A4* and *MAOA* were ligated into the polyclonal site of the pGL3 basic vector (Promega, USA) which contains the promoterless firefly luciferase gene. Initially *BglII* and *HindIII* restriction sites were incorporated into the forward and reverse primers respectively. The core promoters were amplified, according to the initial PCR protocols, and 5µl of PCR product checked by electrophoresis (as previously described). Amplified promoters were then double-digested with *BglII* and *HindIII* restiction enzymes (New England Biolabs, USA) in a 20µl reaction consisting of 15µl of PCR reaction, 5 units of each enzyme, 1 x NEB reaction buffer 2 and brought up to 20µl with distilled water. Reactions were

then incubated at 37°C for 1.5 hours. 1-2μg of pGL3 were similarly double digested. After digestions, reactions were cleaned using a PCR clean-up kit (Qiagen, U.S.A.). 0.5 units of SAP were added to 10μl of digested and cleaned plasmid and incubated at 37°C for 30 mins. before a 15 min deactivation step at 80°C - this removes phosphate groups at exposed sequence ends and avoids self annealing. Promoter fragments were then ligated into pGL3, at a 5:1 molar ratio of insert DNA:vector DNA, with T4 DNA ligase using the Quick Ligation Kit (New England Biolabs, U.S.A.). The constructs were transformed into TOP10 chemically competent cells (Invitrogen, U.S.A.) for storing for future use. Pure plasmid constructs were obtained for experimental use by growing the relevant TOP10 colony in LB media with 50μg/μl carbanomycin at 37°C for 12 hours followed by purification using endotoxin free (EF) midi- and maxiprep kits (Qiagen, USA). To confirm the presence of an insert of the correct size, PCR was performed using primers specific to the pGL3 vector and complementary to sequence surrounding the clonal site, followed by electrophoresis (as described previously). The forward primer is pGL3V_F (5'TGCAAAGGGTGGCTTCTTCA), and the reverse primer is pGL3V_R (5'GCAGTGGGCTCCGTCAGT). The same PCR conditions and protocols were used as for the gene specific primers.

To create a construct with the low frequency SNP allele for TPH2 the above procedure was repeated using the reference sample carrying the required polymorphism. For SLC6A4 and SLC18A2 site-directed mutagenesis was employed. Test vectors were transformed into TOP10 cells for storage, and prepared for transfection from culture using the endotoxin free maxiprep kit (Qiagen, U.S.A.).

Correct sequence insertion was confirmed by amplification, cleaning and sequencing (as described previously) using the primers specific to the pGL3 vector, pGL3V_F and pGL3V_R. pGL3 constructs were purified from a starter culture using a Qiagen miniprep kit, and used as template for the sequencing reaction.

2.2.4.2 Site-Directed Mutagenesis

Site-directed mutagenesis enables mutations to be induced into double stranded DNA for functional testing. The QuikChange® Multi Site-Directed Mutagenesis kit (Stratagene, CA, U.S.A.) was used here which exploits plasmid DNA replication, with the benefit of allowing multiple mutations to be induced simultaneously. Into a 25μl reaction were added 100ng of plasmid DNA, 100ng of each mutagenic primer, 1μl QuikChange® Multi enzyme blend, 0.75μl QuikSolution,

1µl dNTP mix, 2.5µl 10 X QuikChange® Multi reaction buffer and brought up to 25µl with double distilled H_2O. The mutagenic primers are shown in table 2.2.4. The reaction mixture then followed a thermocycling protocol of 95°C for 1 min., followed by 30 cycles of 95°C for 1 min.; 55°C for 1 min.; 65°C for 10 mins., in a Perkin Elmer 9700 GeneAmp® PCR System thermocycler. The reaction mixture was then quickly brought to 37°C, 10 U of *Dpn* I restriction enzyme added which digests parental, nonmutated, methylated DNA,and incubated for 1 hour. 1.5µl of the digested plasmid DNA mixture was then transformed into XL10-Gold® ultracompetent cells, heat pulsed for 30 secs. at 42°C. Mixtures were incubated at 37°C with shaking at 225-250 rpm for 1 hour before 1µl, 10µl and 100µl were grown on agar plates prepared with antibiotic at 37°C for 24 hours.

Table 2.2.4. Site-directed mutagenesis primers used for *SLC6A4* and *SLC18A2*.

Gene	Variant	Primer Name	Primer Sequence (5' to 3')
SLC6A4	SNP1	SERT1mut	TGCCGGCTGCTCCGGGCTCCGCTCCTCCC
	SNP2	SERT2mut	CGCGGCCCCTCCCCGGCGAGCGCAACCC
	SNP3	SERT3mut	CCTGGCGAGCGCAACTCCATCCAGCGGGAGC
	SNP4	SERT4mut	GCCGGCGCCCCTCCACACAGCCAGCGCCG
SLC18A2	SNP2/4	VMATmut2/4	GCGACGGCACGGGCGGGAGGAGGC
	SNP4	VMATmut4	CGACGGCGCGGGCGGGAGGAGGCCG
	SNP5	VMATmut5	CCAGCCCCCCGCCTCCGCTCCCTCCGGC
	SNP6	VMATmut6	CCCCGCTCCCTCCAGCCGTGACGTCAGA

Highlighted nucleotides indicate the minor allele of the respective SNP.

2.2.4.3 Cell Maintenance

Once constructs with correct insert sequence were made they were transfected into the appropriate eukaryotic cell line. Two cell lines were used for this study; HeLa (Helen Lane) cells, a robust cell line useful in optimising transfection experiments, and JAR cells, a human placental choriocarcinoma cell line that constitutively expresses serotonin. These cells were grown and maintained in sterile, clear, 100 mm x 20 mm treated tissue culture dishes (Corning) in growth medium at 37°C in a 5% CO_2 incubator. For HeLa cells growth medium consisted of 1x DMEM (Dulbecco's Modification of Eagle's Medium) without sodium pyruvate (Mediatech, Inc., Herndon, VA, USA), 1% Fetal Bovine Serum and 1x penicillin-streptomycin. For JAR cells growth medium consisted of 1x RPMI 1640 (Roswell Park Memorial Institute) without L-Glutamine (MT), 1% Fetal Bovine Serum, 1x penicillin-

streptomycin and 2 mM L-Glutamine. After 48 hours of growth, cell concentration was reduced by a factor of 1:3 by re-plating into new plates.

2.2.4.4 Transfections

24 hours before transfections, cells (JAR or HeLa) were removed from a culture dish with trypsin/EDTA at 0.25%/0.1%, counted using a haemocytometer and seeded into 96 well, white, clear-bottom tissue culture plates (Corning) at 1 x 10^4 cells per well. After 24 hours the complete growth media was removed from all wells and replaced with 100µl of regular media containing no antibiotic. Into each well was added 50µl of a mixture containing 2µl of pGL3 + test insert in endotoxin free 1 x TE (at 50ng/ul), 0.5µl of pRL-thymidine kinase (Promega, at 50ng/µl) as the internal control, 0.8µl Lipofectamine™ 2000 (Invitrogen) and brought up to 50µl with regular media. The lipofectamine was incubated in media for 5 mins. at room temperature before mixing with the plasmid constructs. After all solutions were added the cells were incubated at 37°C in the CO_2 incubator for 3-4 hours. The regular media was then replaced with 100µl complete media and cells incubated for 36 to 48 hours.

For *SLC18A2*, transfections of two different plasmid DNA preparations were carried out in one 96 well plate, with eight replications for each construct. A second 96 well plate was also made with a different arrangement of construct order. For this gene, therefore, a total of 32 measurements could be obtained for each construct in each cell line. For *SLC6A4* and *TPH2* constructs only one DNA preparation was possible, with 8 transfection replications in two different plate orders, giving a maximum of 16 measurements. Without a second DNA preparation for *SLC6A4* and *TPH2* results are only preliminary.

2.2.4.5 Luminescence Assay

Cells were removed from the incubator and growth media aspirated off. 20µl of 1x Passive Lysis Buffer (Promega) were added to each well and plates agitated on a 4625 Titer Plate Shaker (Lab Line Instruments Inc.) at a frequency setting of '3' for 20 mins. Each plate was then placed in an LD Max Microplate Luminometer (Molecular Devices, USA) where the Dual-Luciferase Assay reagents were added. 100µl of Luciferase Assay Reagent II were added, followed by a 10 second delay, followed by the first (firefly) luminescence reading. 100 ul Stop and Glo reagent were

then added, followed by a 10 second delay, followed by the second luminescence reading. Luminescence measurements were made in relative light units (RLU) and expressed as firefly luciferase / *Renilla* luciferase.

2.2.4.6 Experimental Procedure Controls

As well as the co-transfection of pRL-TK constructs as an internal control for transfection efficiency, other experimental controls were considered. In order to assess levels of luminescence produced in a strongly driven promoter we transfected a pGL3 vector + simian virus 40 promoter construct (Promega, USA). Firefly luminescence measurements were typically 1000 times that of the *SLC18A2* and *TPH2* test constructs, and 100 times the *SLC6A4* constructs. Constructs with low and high copies of the *MAOA* promoter VNTR were contrasted to assess the sensitivity of this experimental procedure to differences in expression. The *MAOA* promoter was used as significant effect of copy number on expression levels has already been documented (Sabol et al., 1998).

2.2.5 Statistical Analysis

2.2.5.1 Nucleotide Diversity

Two measures were used. π is the mean number of pairwise differences between any two haplotypes for a region, divided by the total number of nucleotides. It is equivalent to the probability that any two homologous nucleotides are different and is defined

$$\pi = \frac{\sum_{i=1}^{k} \sum_{j<i} p_i p_j d_{ij}}{L}$$

where d_{ij} is the number of differences between haplotypes i and j, k is the number of haplotypes, p_i is the frequency of haplotype i, p_j is the frequency of haplotype j and L is the length of the haplotype in nucleotides (Tajima, 1983).

θ describes the proportion of nucleotides that are segregating, corrected for the total number of sequences analysed. It is calculated from;

$$\theta = \frac{S / a_1}{\text{length of haplotype in nucleotides}} \quad \text{, where } a_1 = \sum_{i=1}^{n} 1/i \quad .$$

S is the number of segregating sites, i represents the i^{th} sequence and n is the total number of sequences sampled (Tajima, 1989).

2.2.5.2 Activity Measurements

The luminometer generated two luminescence readings: firefly luciferase (FF) and *Renilla* luciferase (Ren). The FF reading represents the test construct, whereas the Ren reading is the internal control for transfection efficiency. Firstly, an average blank reading from a set of wells containing only seeded cells and Dual-Luciferase Assay reagents was taken from the FF and Ren readings. This represents a background luminescence value. If residuals values were obviously non-informative, such as a negative value or value close to zero, then this data point was ignored. Eleven data points were ignored in this way for the *SLC18A2* constructs, out of a total of 160. Three *SLC6A4* datapoints were ignored out of a maximum of 80, and no datapoints were discarded for *TPH2* for which there was also a maximum of 80 transfections. Once the initial FF and Ren readings were corrected for background luminescence, the test reading (FF) was corrected for transfection efficiency by dividing by the Ren value, FF/Ren. To make the variant promoter measurements relative to the most frequent (wild-type) promoter haplotype, a final correction to variant measurements was made by;

$$\frac{\text{variant haplotype FF/Ren}}{\text{most frequent haplotype FF/Ren}} ,$$

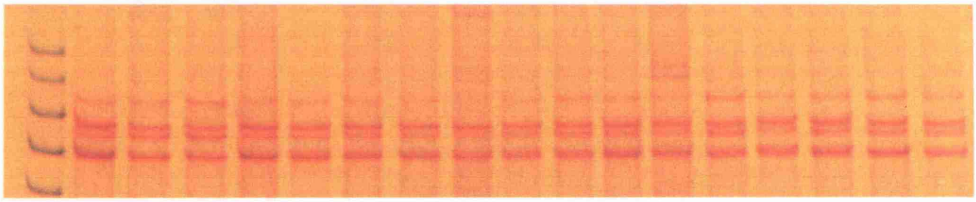
and it is these values that are plotted in figures 2.3.4 to 2.3.6.

2.3 Results

2.3.1 Sequence Diversity of Promoter Regions

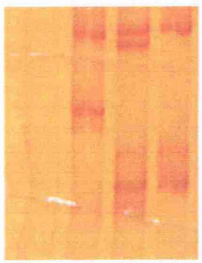
After comprehensively screening the upstream regions of *HTR1A* in a group of Antioquian BPI affecteds for novel variation and typing them for documented SNPs, no variation was detected hence no further analysis was done with this gene. Results from a typical SSCP gel are shown in figure 2.3.1. A negligible level of diversity for this promoter was also found separately by re-sequencing in the Freimer lab at UCLA (Charles Glatt, personal communication).

Figure 2.3.1. SSCP gel for *HTR1A* fragment 3 in BPI affecteds from fourteen Antioquian families and three Spanish controls.



No variation is observed. The non-denatured 100bp DNA ladder is loaded into the first well, diluted 1 in 4. V18, V19 and V39: Spanish controls.

Figure 2.3.2. SSCP results for three Antioquian BPI affecteds at a *COMT* RFLP.



These individuals were found to have three different genotypes: HH, HL and LL respectively. This was used as a control for functionality of the SSCP assay.

After screening the promoters of *TPH2*, *SLC18A2* and *SLC6A4* for novel SNPs by re-sequencing a range of polymorphism was obtained, as summarised in table 2.3.1. Interestingly, SNPs discovered for *SLC18A2* co-occurred on

chromosomes with other SNPs. These allelic associations are represented as haplotypes in table 2.3.2.

Table 2.3.1. Novel SNPs in the promoters of *TPH2*, *SLC6A4* and *SLC18A2* and sequence diversity estimates.

Gene	SNP	Allele	Pos (To T.I.)	Ave. Freq.	AA Freq.	Cauc Freq.	π	θ
TPH2	1	A/G	-301	0.01	0	0.01	0.000035	0.000578
	2	T/A	-477	0.01	0	0.01		
SLC6A4	1	C/G	-277	0.01	0	0.01	0.000244	0.001057
	2	T/C	-177	0.07	0.06	0.07		
	3	C/T	-164	0.01	0	0.01		
	4	G/A	-2	0.01	0.01	0		
SLC18A2	1	G/T	-219	0.01	0.01	0	0.001937	0.001916
	2	G/A	-112	0.01	0.01	0		
	3	G/A	-106	0.10	0.16	0.04		
	4	C/A	-103	0.47	0.61*	0.32		
	5	G/T	-74	0.05	0.06	0.04		
	6	G/A	-62	0.15	0.07	0.23		

Frequencies correspond to the minor allele which is defined by the second nucleotide in the 'Allele' column. * SLC18A2 SNP4 allele 'A' is actually more common than 'C' in African Americans, although 'A' is the minor allele according to the average frequency. AA: African American; Cauc: Caucasian; π, θ: measures of nucleotide diversity; T.I.: transcription intiation site.

Table 2.3.2. *SLC18A2* promoter haplotypes.

SLC18A2 Haplotype	5' 3'	Ave Freq.	AA Freq	Cauc Freq
WT	1 1 1 1 1 1	0.53	0.39	0.67
1	1 1 1 2 1 1	0.17	0.32	0.02
2	1 1 1 2 1 2	0.14	0.07	0.21
3	1 1 1 2 2 1	0.04	0.03	0.06
4	1 1 2 2 1 1	0.10	0.17	0.04
5	1 2 1 2 1 1	0.01	0.01	0
6	2 1 1 1 1 1	0.01	0.01	0

WT: wild-type haplotype defined by the most common allele at each locus, on average, in African Americans and Caucasians; 1: major allele (aggregate); 2: minor allele (aggregate); AA: African American; Cauc: Caucasian.

2.3.2 Functionality of Novel Variation

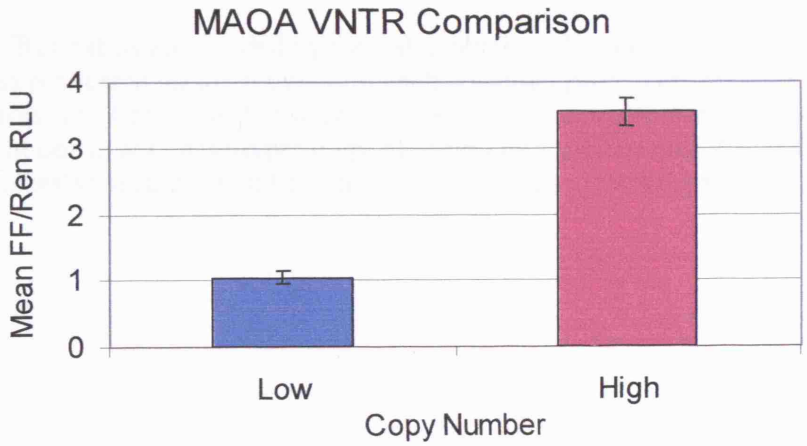
The functionality of these new found SNPs has been tested by cloning the promoter fragments upstream to a luciferase reporter gene in a plasmid vector and co-transfecting with a pRL-thymidine kinase vector into JAR cells. If SNPs are functional then significant differences in luminescence between the major and minor allele constructs are expected. I have focussed on *SLC18A2*, *SLC6A4* and *TPH2*. The results are illustrated in figures 2.3.4, 2.3.5 and 2.3.6. Figure 2.3.3 demonstrates the sensitivity of this test and shows the activities of *MAOA* promoter constructs containing a low copy number (3 copies) and high copy number (4 copies) of a functional VNTR. Previous work has already shown that promoters with four copies of the VNTR transcribe 2 to 10 times more efficiently than those with three copies (Sabol et al., 1998). Results obtained in my assay fall comfortably in this range as the high copy number VNTR transcribed 3.4 times more efficiently than the low copy number.

For both *SLC18A2* and *SLC6A4* variant promoters significantly altered expression. For *SLC18A2* the most significant effects are for haplotypes 3 and 5, both of which demonstrated approximately 0.75 times the expression of the wild-type (unpaired student's t-test, $p<0.01$). Haplotype 1 increased expression by 1.25 times (unpaired student's t-test, $p<0.05$).What's more, the effect for each promoter haplotype is repeated between preps, demonstrating the reliability of these results. In figure 2.3.7, the location of SNPs involved in the haplotypes is illustrated. Haplotype 3 involves SNPs 4 and 5; interestingly, SNP 4 lies in a GC rich region. Haplotype 5 involves SNPs 2 and 4, and these both lie in the same GC rich region. Haplotype one consists solely of SNP 4.

For *SLC6A4* all four SNPs have been shown to increase expression levels (unpaired student's t-test, $p<0.01$); from around 1.6 times (variant 2) to 2.3 times (variant 1) more than the wild-type promoter. It would seem that these SNPs disrupt negative regulation sites in the *SLC6A4* promoter. From figure 2.3.8 it can be seen that SNP 1 lies adjacent to a Sp1 transcription factor binding site, while SNP 4 lies just 2 bp upstream from the transcription start site. The sequence conservation plot between human and mouse reveals that SNPs 1, 3 and 4 lie in non-coding sequence that is more than 50% conserved between the two species (figure 2.3.8).

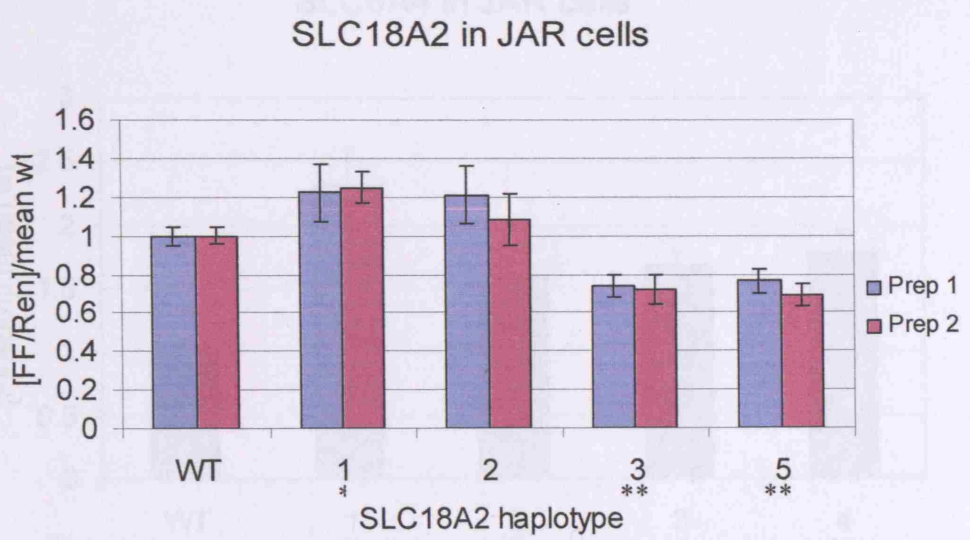
For *SLC18A2* two plasmid DNA preparations were tested, whereas for *SLC6A4* and *TPH2* plasmids were cultured and isolated once. As variability in the DNA isolation procedure can lead to variation in the reporter assay, it is best to test more than one preparation for reliability. The results for *SLC6A4* and *TPH2* therefore need to be repeated before robust conclusions can be drawn, although these preliminary findings are potentially very interesting.

Figure 2.3.3. Comparison of expression levels between *MAOA* promoters containing low and high copies of a functional VNTR.



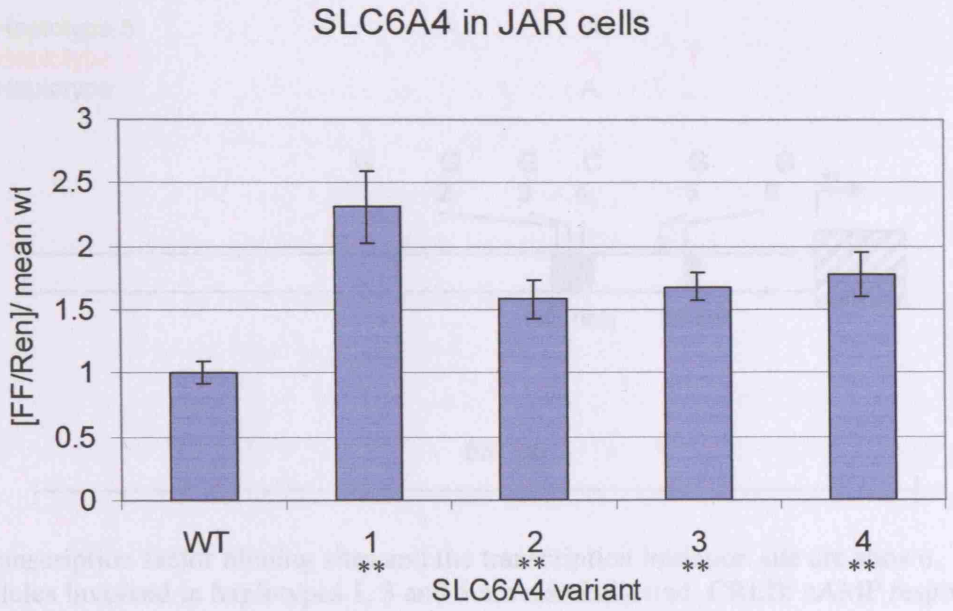
Transfection was replicated 8 times for each construct. Vertical bars represent standard errors. FF: firefly luciferase activity; Ren: *Renilla* luciferase activity; RLU: random light units.

Figure 2.3.4. Mean expression levels in four variant *SLC18A2* promoter haplotypes.



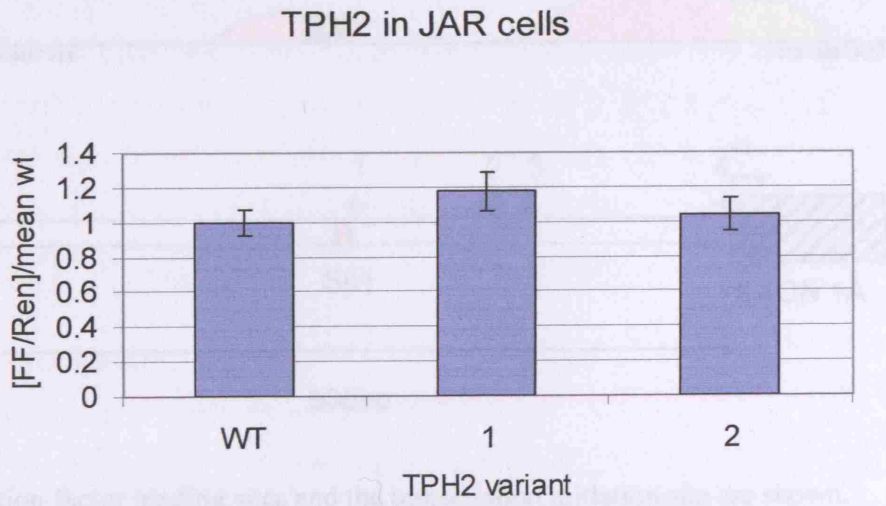
FF/Ren ratios are divided by the value obtained for the WT haplotype. Transfection was replicated up to 16 times for each construct prep. Vertical bars represent standard errors. FF: firefly luciferase activity; Ren: *Renilla* luciferase activity; RLU: random light units; WT: wild-type; Prep: plasmid DNA preparation. * $p<0.05$, ** $p < 0.01$: student's t-test, 2-tailed FF/Ren (prep1 + prep2) vs. wild-type (prep1 + prep2).

Figure 2.3.5. Mean expression levels in four variant *SLC6A4* promoters.



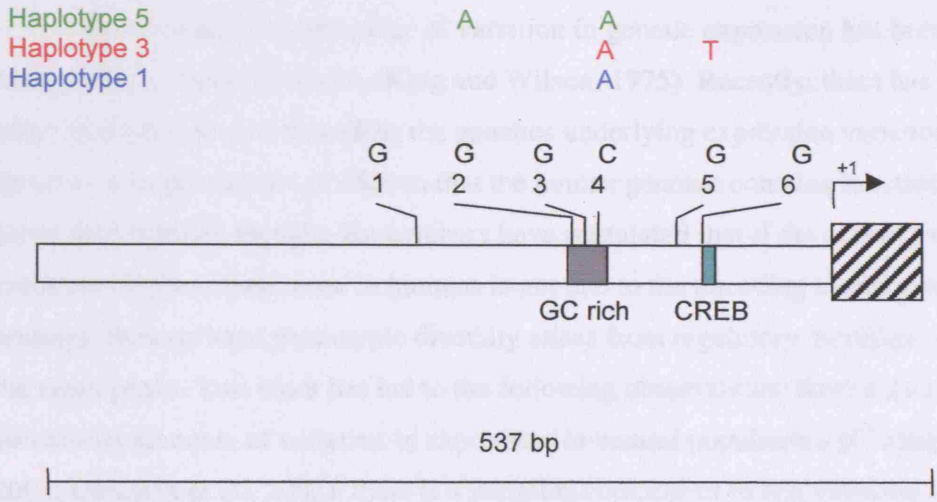
FF/Ren ratios are divided by the value obtained for the most common (wild-type) promoter. Transfection was replicated up to 16 times for each construct. Vertical bars represent standard errors. FF: firefly luciferase activity; Ren: *Renilla* luciferase activity; RLU: random light units; WT: wild-type. ** $p < 0.01$: student's t-test, 2-tailed FF/Ren vs. wild-type.

Figure 2.3.6. Mean expression levels in two variant *TPH2* promoters.



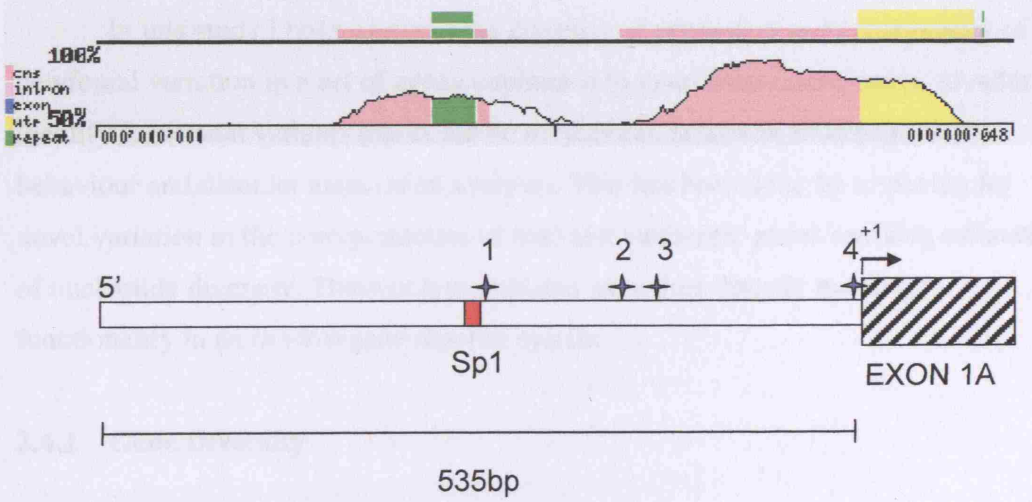
FF/Ren ratios are divided by the value obtained for the most common (wild-type) promoter. Transfection was replicated up to 16 times for each construct. Vertical bars represent standard errors. FF: firefly luciferase activity; Ren: *Renilla* luciferase activity; RLU: random light units; WT: wild-type.

Figure 2.3.7. Location of novel *SLC18A2* promoter polymorphism.



Transcription factor binding sites and the transcription initiation site are shown. Alleles involved in haplotypes 1, 3 and 5 are also indicated. CREB: cAMP response element; +1: transcription initiation site.

Figure 2.3.8. The location of novel *SLC6A4* promoter polymorphism.



Transcription factor binding sites and the transcription initiation site are shown. A conservation plot with mouse sequence is also presented, generated using the rVista software. Sp1: Sp1 transcription factor binding site; +1: transcription initiation site.

2.4 Discussion

The biological significance of variation in genetic expression has been recognised for several decades (King and Wilson, 1975). Recently, there has been a surge of interest in understanding the genetics underlying expression variation, spurred on largely by the revelation that the human genome contains substantially less genes than initially thought. Researchers have postulated that if the expansive spectrum of phenotypes seen in humans is not due to the encoding of many unique proteins, then perhaps phenotypic diversity arises from regulatory modifications of the same genes. This work has led to the following observations: there exists substantial amounts of variation in expression in natural populations (Cheung et al., 2003; Oleksiak et al., 2002); there is a heritable component to this variation (Cheung et al., 2003; Yan et al., 2002); *cis*-acting variants have been shown to have functional effects (Hoogendoorn et al., 2003; Yan et al., 2002); and about 10% of human promoters have functional polymorphisms (Hoogendoorn et al., 2003). Together, these lines of evidence provide sound support for the substantial biological influence of promoter variants, and for promoters as targets of natural selection. In particular, this variation may be influential in quantitative and complex phenotypes; traits that are reliant on levels of protein and hormone.

In this study I have assessed the diversity of promoters and functionality of newfound variation in a set of genes implicated in neurobehavioural traits, in order to identify functional variants that could be used as candidates in neuropsychiatric behaviour and disorder association analyses. This has been done by screening for novel variation in the core promoters of four serotoninergic genes enabling estimates of nucleotide diversity. These polymorphisms were then directly tested for functionality in an *in vitro* gene reporter system.

2.4.1 Gene Diversity

Novel variation was found in the three pre-synaptic genes: *TPH2*, *SLC6A4* and *SLC18A2*. For the *HTR1A* serotonin receptor no variation was discovered. Separately at UCLA, a very low level of variation was also revealed, confirming that the *HTR1A* core promoter shows very little variation. This suggests strong preservation of the sequence immediately upstream of the *HTR1A* coding DNA. The HTR1A receptor is

found both presynaptically and postsynaptically and inhibits the firing rate of serotonergic neurons, synaptic release of serotonin and downstream serotonergic gene activation (Barnes and Sharp, 1999). As variation in the amount of this gene product can influence several functions of serotonin a strong selective constraint on expression levels may have been enforced as a result.

The nucleotide diversity (θ) values calculated here show the *SLC18A2* promoter (1.916×10^{-3}) to be approximately 10 times more diverse than its coding regions (2.22×10^{-4}), and the *SLC6A4* promoter (1.057×10^{-3}) to be approximately 3 times as diverse as its coding DNA (3.44×10^{-4}) (Glatt et al., 2001). The differences are substantially increased when only non-synonymous coding SNPs are considered. These results show that in humans much larger amounts of variation persist in the promoters of these two genes than coding regions.

Of the three pre-synaptic genes the nucleotide diversity level for *TPH2* was shown to be the lowest, with a θ value of 5.78×10^{-4}. This is more than the θ value for *TPH2* coding sequence (4.6×10^{-4}), but less than for intronic sequence (7.28×10^{-4}), as reported in a previous study (Breidenthal et al., 2004). The fact that this promoter is not as diverse as intronic sequence, and is the most conserved of the pre-synaptic genes tested here, suggests conservation of this promoter sequence. This is further supported by the fact that *TPH1*, a close functional homologue, catalyses the rate-limiting step in serotonin synthesis in the periphery (Lesch, 2001). The importance of a functional constraint on the regulation of this gene is therefore implied.

The average θ for the promoters of the three presynaptic genes is 1.184×10^{-3}, which is greater than the average for their non-coding intronic regions at 5.47×10^{-4}, and the coding regions at 3.42×10^{-4} (Breidenthal et al., 2004; Glatt et al., 2001; Glatt et al., 2004). These estimates are also substantially larger than genome-wide diversity averages, with estimates of 5.43×10^{-4} (θ) for coding DNA, 5.3×10^{-4} (θ) for non-coding intragenic DNA (Cargill et al., 1999), 8.8×10^{-4} (π) for non-coding intergenic DNA (Yu et al., 2002) and 7.571×10^{-4} (θ) as a genome-wide average (Marth et al., 2003). Additionally, these pre-synaptic serotoninergic promoters show an aggregate θ more than twice that for a randomly selected group of promoters (Hoogendoorn et al., 2003). Substantial promoter polymorphism has therefore been revealed here, not only compared to the coding genome but also to other promoters, and is largely reflective of the high diversity seen in *SLC18A2* and *SLC6A4*. If this variation affects promoter function a spectrum of expression levels may arise in human populations, which may

lead to a spectrum of behavioural phenotypes. It is important, therefore, to determine whether the novel variants discovered here are functional.

2.4.2 Variant Functionality

Two of the three (67%) genes with cloned polymorphic promoters, *SLC18A2* and *SLC6A4*, showed statistically significant functional differences between their promoter haplotypes. This is a higher ratio than for estimates of the number of genes possessing functionally variant promoters in two previous studies. In a group of genome-wide randomly selected genes thirteen out of thirty-eight genes (34%) had promoters with functionally variant haplotypes (Hoogendoorn et al., 2003). A similar ratio of eight out of twenty-five (32%) was found for a group of genes previously shown to be differentially expressed in the brains of schizophrenia patients (Buckland et al., 2004).

Of the four variant *SLC18A2* haplotypes tested for functional differences two, haplotypes 3 and 5, showed strong decreases compared to the most common haplotype ($p<0.01$) and one, haplotype 1, showed a modest increase ($p<0.05$); therefore three out of four variant haplotypes are functional for *SLC18A2*. These haplotypes share one SNP that is present in a GC rich region and both of haplotype 5's variants are in this GC rich sequence. CpG islands are associated with gene regulation; methylation of CpG islands is known to repress expression and has been associated with human disease (Abdolmaleky et al., 2005; Robertson and Wolffe, 2000). Additionally, mutations found in GC rich regions have previously been shown to have a negative effect on gene expression (Maniatis et al., 1987). *SLC18A2* mediates the transport of neurotransmitters such as serotonin, dopamine and norepinephrine into neuronal synapses via cellular vesicles. A restriction of its function has been shown to directly alter levels in the brains of mice (Takahashi et al., 1997) and detrimentally affect murine response to neurological drugs and toxins (Uhl et al., 2000). Variant haplotypes that alter the expression of this gene, as identified here, may therefore be good functional candidates for common human disorders with a neuronal basis. The effect of individual SNPs on expression appears complex, however, as although haplotypes 3 and 5 both contain SNP 4 and both show reduced expression, haplotype 1 consists solely of SNP 4 and shows increased expression.

All four of the *SLC6A4* variant haplotypes tested for functionality were composed of a single SNP, and all four conferred an increase in expression compared to the most common haplotype at a significance level of $p<0.01$; suggestive that down-regulation mechanisms have been disrupted. Haplotype 1 showed the largest increase with more than twice the activity compared to the most common form. Interestingly, SNP 1 lies in sequence that is more than 50% conserved with murine sequence and is adjacent to an Sp1 transcription factor binding site (figure 2.3.8). This suggests that this sequence has been important to the *SLC6A4* gene before the primate and rodent lineages diverged, indicative of functional conservation where novel mutation is more likely to have a deleterious effect. The serotonin transporter is responsible for the re-uptake of synaptic serotonin; a neurotransmitter involved in many functions including CNS growth and development, pain reception, behaviour and mood (Lucki, 1998). As it has a direct effect on serotonin functionality, *SLC6A4* has been a target gene for researchers interested in behaviour and mood. Extensive research has associated this gene with a wide range of psychiatric disorders and human behaviours including; attention deficit hyperactivity disorder (Curran et al., 2005), depression (Willeit et al., 2003), bipolar affective disorder (Mynett-Johnson et al., 2000), obsessive-compulsive disorder (Ozaki et al., 2003), and a role in response to fearful facial expressions has also been implied (Hariri et al., 2002). Furthermore, a functional promoter polymorphism known to significantly alter gene expression levels (the 5-HT transporter length polymorphic region) has repeatedly been implicated in these disorders and behaviours. Therefore, the potential importance of the functional promoter variants identified here is highlighted. For all haplotypes an upregulation of *SLC6A4* is indicated thereby implying an increase in re-uptake activity and a decrease in synaptic serotonin.

A relatively large amount of the variation discovered here has proved to be functional: in all, ten uncommon variant promoter haplotypes were tested for functionality, and seven of these were shown to alter expression levels compared to a more common haplotype; a higher percentage than in two other studies (Buckland et al., 2004; Hoogendoorn et al., 2003). This is perhaps suggestive of the biological significance of the sequence variants and the promoters in which they are contained. However, in a multi-staged functional assay, as was conducted here, there are several stages in which experimental error and procedures may influence results. For example, absolute luminescence values were fairly small in these assays and may lead

to overestimations of the significance of differences, whether they are real biological changes or random stochastic ones. The importance of repeating results is consequently increased, and effects were repeated for *SLC18A2* constructs. Unfortunately, due to time and economic restrictions, not all *SLC18A2* haplotypes were able to be tested in a functional assay. The same restrictions apply for second plasmid preparations of *SLC6A4* and *TPH2* haplotypes, which are necessary for confirmation of findings presented here.

A large amount of sequence diversity has been demonstrated in the promoters of *SLC6A4* and *SLC18A2*, a large proportion of the genes with polymorphic promoters showed functional variation between haplotypes (67%) and a large proportion of the cloned variant haplotypes demonstrated functional differences (70%). Considering the functionality results to be reliable, a question of interest raised here is why high levels of functional polymorphism persist in natural populations? One answer may be the effects generated by the polymorphic promoters, via differential expression rates, are not significantly detrimental to be acted upon by selection. Regulation of expression in these genes is not likely to be solely dependent on *cis*-acting promoters. Rather, more distal or *trans*-acting factors may be more influential. Work done by Morley *et al* (2004) suggests that eukaryotic gene expression is likely controlled by major *trans*-acting loci (Morley et al., 2004), and the majority of heritable variation in expression has been shown to be *trans*-acting in yeast and mice (Schadt et al., 2003; Yvert et al., 2003). The effect of functional variants in the promoters of *SLC6A4* and *SLC18A2* on their expression may only be moderate as a result. The phenotypic consequence of such variation may not be severe enough to be removed from a population by selective forces; it may only alter behavioural phenotypes slightly and in a way that does not reduce an individual's reproductive success. When these variants interact opportunistically with other genetic determinants or environmental factors then adverse phenotypic effects may ensue. For example, as *SLC6A4* has been implicated in mood, promoter haplotypes that moderately reduce the amount of synaptic serotonin may make someone more depressed than average, but still not clinically depressed. Major depression would only follow if, say, an appropriate mutation was to occur on another mood related gene or the appropriate environmental event was encountered.

This work has revealed novel polymorphisms in the promoters of some pre-synaptic serotoninergic genes, at relatively high frequencies. Furthermore, I have demonstrated functionality for a large proportion of this variation; reliably for *SLC18A2* and tentatively for *SLC6A4*. The fact that three of the variant *SLC6A4* haplotypes (defined by a single SNP) and two of the variant *SLC18A2* haplotypes occur at combined frequencies of less than 0.05 should not undermine potential biological significance; complex traits may have a heterogeneous genetic architecture, where many low frequency alleles give rise to a phenotype.

An interesting extension to this work would be to include more genes involved in serotonin transmission. For example, a total of fourteen unique human serotonin receptors exist, grouped together into seven subfamilies by pharmacological function and structure (Barnes and Sharp, 1999). Inclusion of more genes in studies of this kind could further highlight which of the many serotoninergic genes should get priority as candidate genes in association studies. The method employed here for detecting new variants has likely left some undiscovered as 96 chromosomes from one ethnic group does not provide much power to detect SNPs of very low frequency (Glatt et al., 2001); such SNPs may also be functional. Obviously ethnic groups other than African Americans and Caucasian Americans may also harbour functional promoter variants.

Nonetheless, the functionally relevant *cis*-acting promoter variants in *SLC18A2* and *SLC6A4* discovered here represent excellent candidate loci for association analyses of neurobehavioural traits. What's more, as these variants are likely to have relevant functional consequences, positive association studies using these variants will increase understanding of the molecular basis of these traits.

CHAPTER 3: THE GENETICS AND EVOLUTIONARY HISTORY OF THE ANTIOQUIA POPULATION ISOLATE

CHAPTER 3: THE GENETICS OF THE ANTIOQUIA POPULATION ISOLATE AND ITS EVOLUTIONARY HISTORY

3.1 Introduction

Since the New World was successfully colonised by modern humans 15 to 20 thousand years ago, the most drastic demographic change occurred after the arrival of the first Europeans on American lands in the late 15th century. For several centuries Europe battled for control of the fruitful, newly discovered lands of the Americas. By the time independence had been declared from the European nations in the late 18th and early 19th centuries, and the European colonisers returned to their home countries, the Native American populations had been largely replaced with people of mixed Amerindian, European and African descent. An example of such an admixed population founded during the colonial period is Antioquia from northwest Colombia.

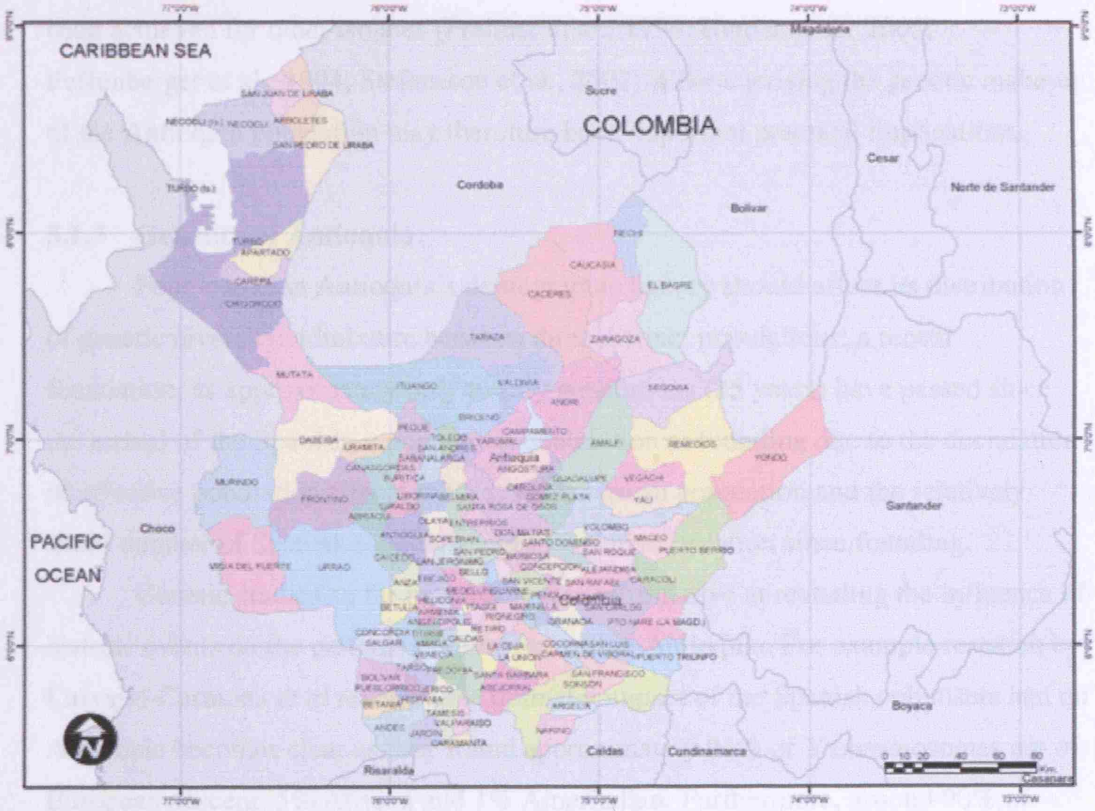
3.1.1 Antioquian History

Antioquia is a mountainous Colombian province situated in the North-West of the country covering approximately 64,000 squared kilometres. Containing the West Andes mountain range, to the north it stretches to the Caribbean coastline.
Due to its location, Antioquia would have been a gateway to other South American lands encountered by the first modern humans, and archaeological evidence indicates colonisation first occurred around 8-10 KYA (Cooke, 1998; de Jaramillo and Vega, 2001); genetic studies are in general agreement with this figure (Bortolini et al., 2003; Ruiz-Linares et al., 1999). Although relatively little is known about these first settlers, it is agreed that the mountainous terrain would have separated them into small, isolated hunter-gatherer populations with little population mixing. This is seen in the low genetic diversity, both mtDNA and autosomal, of the native Amerindian populations of this region (Barrantes et al., 1990; Bortolini et al., 2003), as well as the high levels of genetic structure between them (Mesa et al., 2000). Mesa *et al* (2000) showed, using autosomal loci, the level of genetic structure to be around 6.8% between Colombian Amerind populations, a figure similar to other native South American regions but twice that of other continents (Cavalli-Sforza, 1994).

The Antioquian population was drastically influenced by the arrival of the Spanish in 1500 AD, culminating in a large decimation of indigenous people. By the

17th century, post-Columbian Antioquia consisted of an admixed population of European-Amerinds, African-Amerinds (resulting form the African slave trade), and a small number of non-admixed individuals. After independence from Spain had been gained in 1819 Antioquia was a fairly small, isolated population left to develop on her own; and a flourishing coffee industry gradually developed. As a result, Antioquia experienced a population explosion in the late 19th and early 20th centuries, helped by a small amount of immigration from other Colombian regions.

Figure 3.1.1. Map of Antioquia and its municipalities. (Kindly provided by Dr. Andres Ruiz-Linares.)



3.1.2 Contemporary Antioquia

Currently, Antioquia supports a population of around 4.7 million people, a large proportion of whom live in the capital Medellin situated in the Aburra valley (de Jaramillo and Vega, 2001). The majority of Antioquia is admixed with mostly Amerindian and European ancestry (98.5%), lower Amerindian-African ancestry

(approx. 1%) and a small non-admixed Amerindian component (approx. 0.25%) (de Jaramillo and Vega, 2001). Minimal immigration has occurred in the province since its founding.

Due to its history and geographic location, Antioquia can be regarded as a geographic population isolate, with a unique demographic history. Population isolates are special in the sense that, as a result of founder effects and small effective population sizes, they may have increased genetic homogeneity, decreased disease and trait heterogeneity and extensive linkage disequilibrium. As a consequence, these special populations may have increased power to detect genes of multifactorial traits and disorders which are likely to be influenced by several genes each of modest effect (Escamilla, 2001; Varilo and Peltonen, 2004; Wright et al., 1999), and successes have been achieved for other isolates (Freimer et al., 1996; Hamet et al., 2005; Puffenberger et al., 1994; Stefansson et al., 2002). Characterising the genetic make-up of the Antioquia population may therefore have important practical implications.

3.1.3 Genetics of Antioquia

Four events in Antioquia's demographic history should affect its distribution of genetic diversity: admixture between three distinct populations; a recent foundation, as approximately only twenty generations (25 years) have passed since the arrival of the Spanish; a population contraction at founding due to the decimation of effective population size of the Native American population and the relatively small number of Spanish founders; and the relative isolation since founding.

Genetic studies so far have been very informative at revealing the influence of historic events on the patterns of diversity within Antioquia. For example research by Carvajal-Carmona *et al* revealed the dramatic impact of the Spanish colonisers had on Antioquia becomes clear as they found approximately 94% of Y chromosomes are of European descent, 5% African and 1% Amerindian. Furthermore, around 90% of mtDNA is Amerindian (Carvajal-Carmona et al., 2000). This suggests that the majority of Antioquia was founded by Spanish males and Amerindian females, consistent with male Amerind genocide. When combined with blood-group data which apportioned the population into 70% European, 15% Amerind and 15% African (Sandoval et al., 1993), these findings suggest sex-biased admixture in Antioquia, in favour of Spanish males (Carvajal-Carmona et al., 2000; Sandoval et al., 1993). Interestingly, similar trends have been obtained for other Latin American

populations such as the Central Valley of Costa Rica, Mexican American and various Brazilian populations (Bortolini et al., 1999; Carvajal-Carmona et al., 2003; Green et al., 2000; Merriwether et al., 1997).

Carvajal-Carmona *et al* (2003) studied the genetic diversity within Antioquia in more detail. They used three genetic systems (Y chromosome, mtDNA and autosomal microsatellites) to make assessments of population structure, genetic diversity and LD in Antioquia, a Latin-American population isolate from the Central Valley of Costa Rica (CVCR), and a well known European population isolate from Finland, the Saami. They revealed a close genetic identity of Antioquia and the CVCR. This was explained by the observation that both populations have undergone admixture with Spanish male colonisers from the same region of Spain, and the Native American component of both populations are from the Chibchan-Paezan linguistic group; so both populations have relatively recent, common, ancestral DNA. Y chromosome diversity was found to be higher in the Americans than the Saami, consistent with theories of directional mating; while mtDNA diversity was lower, explained by high population structure of pre-Columbian Native Americans, as well as from a reduction in Ne of female Amerinds as a result of the Spanish arrival. Using nineteen genome-wide markers, the autosomal diversity was similar in the New World populations and the Finns.

3.1.4 Population Linkage Disequilibrium

Another important measure of genetic diversity within populations is the extent of genomic linkage disequilibrium (LD). Linkage disequilibrium describes the extent of allelic association within a genome and can reflect population demographic history. For example, small populations of constant size theoretically possess extensive LD whereas in large, rapidly growing populations LD should be short-range (Laan and Paabo, 1997; Slatkin, 1994; Wright et al., 1999). Admixed populations may also show extensive LD, but this is dependent on the frequencies of alleles in parental populations and to some extent the relative contributions of each parental population (Patterson et al., 2004). Knowledge of population LD patterns is important for indirect association studies which rely on association between a disease allele and a marker allele.

A thorough analysis of genome-wide LD in populations with different demographic histories was performed by Reich and colleagues (2001). These authors

concentrated on bi-allelic markers, and were concerned with estimating strength of LD at various distances (5, 10, 20, 40, 80 and 160kb) from a core SNP centred in a gene, and looked at these patterns in nineteen genome-wide genes. Using a population of northern European descent they demonstrated: a good correlation exists between LD strength and physical distance; that the half-length of LD (distance at which D' falls below 0.5) is approximately 60kb; and substantial variation exists in strength of LD between different genomic regions. When they repeated this analysis in a population from West Africa a marked decline in LD was observed, as half-life was reduced to less than 5kb in this population. Using simulations, the authors concluded the best explanation for the high LD in northern Europe is a population bottleneck that occurred 27,000 to 53,000 years ago; although they were not able to determine if this bottleneck was specific to northern Europe or common to many non-African populations. Their results also have implications for gene mapping, as they showed that an average genome level of LD may not be a useful parameter in these studies since substantial genome-wide variation was revealed. Additionally, as population-specific LD patterns were demonstrated some populations may be better suited for long-range mapping, such as the European populations that harboured longer-range LD. Other populations with shorter-range LD, including those from Africa, may be more useful in fine mapping studies.

Carvajal-Carmona *et al* (2003) assessed the extent of LD in Antioquia using nine pairs of randomly distributed autosomal STR markers, with each pair representing a pre-determined distance (Carvajal-Carmona et al., 2003). A major finding of this study was that short-range LD was shared between the two populations, suggesting a shared ancestry. The mean distance in significant LD for Antioquia was shown to be 0.43cM, which was significantly less than in CVCR at 1.9 cM. This suggests that Antioquia does not harbour as strong LD as CVCR, possibly because of larger founder Ne sizes in Antioquia or greater expansion of populations since the founding event. Alternatively, as the markers used in this study were selected on the basis of high LD status in the CVCR population (discovered in an earlier study), higher LD in CVCR could represent marker ascertainment bias. To achieve a more comprehensive and objective estimate of LD strength in Antioquia more, randomly selected, genome-wide marker loci could be used. Comparisons to populations with alternative demographic histories would be useful in determining the relative extent of LD in Antioquia; due to its demographic history substantially

stronger LD would be expected in Antioquia compared to old, large, outbred populations.

3.1.5 This Study

The genetic studies of Antioquia to date, although informative, have several limitations as to how well they describe population genetic diversity. Firstly, although mtDNA and the NRY (non-recombining Y chromosome) have been analysed comprehensively by previous studies (Carvajal-Carmona et al., 2003; Carvajal-Carmona et al., 2000), less attention has been paid to autosomal genetic systems. mtDNA and the NRY have been employed extensively in evolutionary and population genetic studies, as molecular variation is generated by stochastic and mutation events alone with no recombination to compound interpretation. There are restrictions to the amount of information they can provide, however, such as: they each represent only one locus; they have an effective populations size ¼ that of autosomal loci, and therefore more prone to stochastic variation; due to being uniparentally inherited, they reflect genealogies of either males or females and so may not be representative of entire populations (Seielstad et al., 1998). In Carvajal *et al*'s study (2003) each of the autosomal markers does represent a separate locus; however, only one marker in each of nine genomic regions was assessed and so values are prone to stochastic errors. To get a more complete, unbiased picture of the autosomal genetic structure many randomly selected markers at various genomic regions are required, so that mean values can be estimated for a region.

Secondly, a random and numerous marker set should also be used for estimates of LD, which have so far been made using a limited set of microsatellite loci that may be prone to an ascertainment bias (Carvajal-Carmona et al., 2003). Furthermore, estimates based on bi-allelic markers are likely to be more useful than estimates from multi-allelic systems for indirect association studies; due to their genome-wide abundance, continual expansions in their discovery and development of efficient genotyping technologies, SNPs are increasingly the markers of choice for geneticists.

Finally, the study by Carvajal *et al* (2003) was limited to Antioquia and two other population isolates. It may also be useful to contrast the genetic diversity within Antioquia to populations with a range of demographic statuses, to further illuminate the 'special' nature of this population.

Here, to further explore the genetics of Antioquia at the autosomal level I have used a large, bi-allelic dataset (425 SNPs) from seventeen autosomal regions of functional importance, closely related to the dataset used by Reich *et al* (2001). Included in analyses are four parental populations with unique and interesting demographies: the West African Beni (from Nigeria); a Spanish population; and two Native American populations, the Chipewayan population from Canada (a member of the NaDene linguistic group) and the Ticuna from Colombia. Genetic diversity and extent of LD within these populations will be estimated, and the genetic relatedness to one another investigated. Diversity is expected to be high in Beni and low in Ticuna and Chipewayan, due to their respective demographic statuses: Beni represents an old, outbred population with a large ancestral effective population size; whereas the Native Americans are small, isolated populations whose ancestors are likely to have passed through at least one population bottleneck (Excoffier and Schneider, 1999; Mesa et al., 2000). Diversity in Antioquia may be high as a result of historical admixture, especially if the parental populations are genetically distant. Similarly, a close genetic affinity is expected between Antioquia and Spain, and the relatedness at the autosomal level will be investigated, as it will between Antioquia and two other potentially ancestral populations; the Beni and Ticuna.

The amount of data on the extent of LD in Antioquia is limited and, as an admixed population, levels could be extremely high (Collins-Schramm et al., 2003). Since seventeen unlinked genomic regions are represented here by up to 35 SNPs spanning around 160kb per region, a comprehensive assessment of LD can be made. As populations from four different continents are included, this study also provides an opportunity to compare LD levels between several populations with diverse demographic histories. If demographic history is the primary factor that influences LD, as some studies suggest it is (Laan and Paabo, 1997; Reich et al., 2001) distinct LD patterns would be expected between these study populations.

Observations made here on diversity, population structure and strength of population LD will describe in more detail than previous research, the genetic structure of Antioquia at the autosomal level. The nature of Antioquian admixture will be further exposed and, as some of the study populations are potentially ancestral to Antioquia, this work may also illuminate Antioquia's evolutionary history. Not only will this information be useful in defining the genetic identity of Antioquia, but could also influence genetic association analyses of complex traits in Antioquia.

3.2 Methods

3.2.1 Population samples

I analysed genetic data from five population samples: Antioquia (n=63); the North Amerindian Chipewayan (n=43); the South Amerindian Ticuna (n=48); a Spanish sample (n=98); and Beni (n=36). These populations represent South America (Antioquia and Ticuna), North America (NaDene), Europe (Spanish), and West Africa (Beni).

Collection of this dataset was carried out in collaboration with the laboratories of Dr.'s Damian Labuda and David Reich. Dr. Labuda contributed the Chipewayan sample and Dr. Reich contributed the Beni sample. DNA was extracted from the Antioquians, Spanish and Ticuna in Dr. Andres Ruiz-Linares' lab.

3.2.2 SNP Discovery and Genotyping

SNP discovery and genotyping were carried out in the lab of Dr. Reich (Reich et al., 2001). In summary, the method of SNP detection follows two main steps. Initially, identification of core SNPs entailed screening more than 3000 genes in a multi-ethnic panel. Core SNPs needed to satisfy two citeria: at least 160 kb of sequence known in one direction from the SNP, and an average minor allele frequency of at least 35% in the multi-ethnic panel. Second, novel variants were discovered at 5, 10, 15, 20, 40, 80 and 160 kb from the core SNP, by initially sequencing around 2 kb at each of the distance 'bins' in 44 individuals (88 chromosomes) of a Caucasian American population. Indentification and genotyping of novel 'bin' SNPs was automated using PolyPhred (Nickerson et al., 1997), and selected if they adhered to Hardy-Weinberg equilibrium and had minor allele frequencies of at least 15/88. Later, 141 individuals from diverse ethnic backgrounds (including 2 non-human primates) were also screened for SNPs at the various distance bins. The study populations were then genotyped for the selected SNPs using single-base extension and one of three different kinds of detection methods: mass spectroscopy, fluorescence polarisation or sequencing gel. More detail on SNP discovery can be found in Reich *et al* (2001).

3.2.3 Study SNPs

For this study genotypic data from seventeen genomic regions was used. Information on these seventeen regions is summarised in table 3.2.1. From these datasets certain SNPs were removed, dependent on the analyses to be carried out.

Table 3.2.1. Genomic regions and numbers of SNPs used in the study.

Gene Symbol	Chromosome	Size (bp)	# SNPs
ACVR2B	3	160467	18
TGFB	5	160861	28
DDR1	6	160050	28
GTF2H4	6	159018	27
COL11A2	6	39183	14
LAMB1	7	158703	22
WASL	7	76585	19
SLC6A12	12	160927	37
KCNA1	12	162077	23
SLC6A3	12	159400	22
PCI	14	159252	36
PRKCBI	16	251652	25
SCYA2	17	162407	35
NF1	17	80937	23
PA12	18	160900	35
IL17R	22	160316	15
*HCF2**	22	160778	18
	TOTAL	2533513	425

SNPs: total number of SNPs that were genotyped for a gene. *The name *HCF2* has been withdrawn by the Human Genome Nomenclature Committee (HGNC) and replaced with *SERPIND1*.

3.2.3.1 Inter-Population Analysis

SNPs for which genotyping was unsuccessful in any one population were removed from all populations. The remaining markers represent subset 1. Nucleotide diversity, π (equivalent to average gene diversity over all loci)(Nei, 1987; Tajima, 1983), was calculated and compared between populations. Patterns of genetic structure within the five populations was assessed by computing pairwise F_{ST}s, estimated by a fixation index.

3.2.3.2 Linkage Disequilibrium

Pairwise LD was assessed for distances spanning approximately 5, 10, 20, 40, 80 and 160 kb. The SNP with the most frequent minor allele (over all populations) at each distance bin from the core SNP was selected for analysis. Any SNP that was monomorphic in any one population was removed from all analyses. This represents SNP subset 2. Population sample sizes were made equal for each gene region.

3.2.3.3 ML Haplotypes

ML (maximum likelihood) haplotypes were estimated for the four regions that showed the strongest degree of LD (*DDR1*, *WASL*, *NF1*, *HCF2*). SNPs were removed from the original dataset if heterozygosities were less than 5%, and if there was more than 10 % missing data, across all populations. All individuals still with missing data were removed from the analysis.

3.2.4 Statistical Analyses

Most analyses were carried out using Arlequin version 2.000 (Schneider et al., 2000) (http://lgb.unige.ch/arlequin/), Powermarker version 3.09 (Liu and Muse, 2005) Liu (Liu and Muse) (http://www.powermarker.net) and Phylip version 3.6 (Felsenstein, 2004) (http://evolution.genetics.washington.edu/phylip/phylipweb.html).

3.2.4.1 Gene diversity

In Arlequin v2.000 the π measure of nucleotide diversity was calculated, described as the average gene diversity over all loci, estimated from;

$$\hat{\pi}_n = \frac{\sum_{i=1}^{k} \sum_{j<i} p_i p_j \hat{d}_{ij}}{L}$$

where $\hat{d}_{ij}$ is an estimate of the number of mutations between haplotypes i and j, k is the number of haplotypes, p_i is the frequency of haplotype I, p_j is the frequency of haplotype j and L is the total number of loci. For this analysis haplotypes are assumed from genotypic data (Nei, 1987; Tajima, 1983).

3.2.4.2 Genetic Structure

Pairwise F_{ST}s were used to determine genetic structure of the populations using an AMOVA (analysis of molecular variance) from which a ratio of covariance values at different heirarchial levels is used to generate F_{ST};

Source of Variation	Degrees of freedom	Sum of squares	Expected mean squares
Among populations	P-1	SS (AP)	$n\sigma_a^2 + \sigma_b^2$
Within populations	2N-P	SS (WP)	σ_b^2
Total	2N-1	SS (T)	σ_T^2

where σ_b^2 is the covariance within a population and σ_a^2 is the covariance between populations. The F_{ST} is calculated from;

$$F_{ST}=\frac{\sigma_a^2}{\sigma_T^2}$$

and is equivalent to a comparison of the probability that two randomly drawn alleles are identical by descent within a population to the probability of IBD between two populations (Excoffier, 2000; Rousset, 2000; Weir, 1996).

A neighbour-joining tree for the five populations was generated considering the pairwise F_{ST}s as genetic distances, using Phylip v3.6.

3.2.4.3 Linkage Disequilibrium

Statistical significance of allelic associations was determined by estimating Fisher's exact test p values between locus pairs using Arlequin v2.0. For genotypic data with unknown phase this program implements a likelihood-ratio test to generate a test statistic, S, derived from the formula

$$S = -2\log(\frac{L_{H^*}}{L_H})$$

where L_{H^*} is the likelihood of the data assuming linkage equilibrium and calculated from the product of the allele frequencies, and L_H is the likelihood of the data not assuming linkage equilibrium calculated by estimating maximum likelihood haplotypes using the expectation maximisation algorithm (Excoffier and Slatkin, 1998; Slatkin and Excoffier, 1996). This test statistic follows a Chi square distribution with $(k_1\text{-}1)(k_2\text{-}1)$ degrees of freedom where k_n is the number of alleles at the nth locus. The empirical distribution of L_H is calculated from;

i. Permute the genotypes between individuals at one locus.

ii. Re-estimate L_H by the EM algorithm. L_{H^*} is not affected by this permutation.

iii. Repeat steps i. and ii. a large number of times to get the null distribution of L_H and therefore S.

The qualitative measure of LD, Lewontin's D', was estimated in PowerMarker v3.09.

3.2.4.5 Haplotype Analysis

Maximum likelihood (ML) haplotypes were estimated using the expectation maximisation algorithm in PowerMarker v3.09. PowerMarker describes this as the process of estimating haplotypes that maximise the likelihood of genotypes, described by;

$$L(F) = \prod_{i=1}^{n} \Pr(G_i \mid F)$$

where F is the haplotype frequency, G_i is the genotype of the i^{th} individual and n is the sample size. The expectation step estimates genotype frequencies using prospective haplotype frequencies;

$$P_i = \sum_{[j, j']S_i} p_j p_{j'}$$

where P_i is the frequency of genotype G_i, [j, j'] S_i is the ordered pair of the jth and j'th haplotypes for genotype G_i, and p_j and $p_{j'}$ are current frequencies of the jth and j'th haplotypes. S_i is the set of ordered pairs that constitute the genotype G_i.

The maximisation step involves calculating the haplotype frequencies based on the genotype frequency calculated in the previous expectation step. It is described by;

$$p_k = \frac{1}{2n} \sum_{i=1}^{n} \sum_{[j, j']S_i} \frac{m_{ij,} p_{i,} p_{j'}}{P_i} ,$$

where k is the number of possible haplotypes and $m_{jj'}$ is 2 if j = j' = k (i.e. homozygous for a haplotype), 1 if $j \neq j'$, j = k or j' = k (i.e. heterozygous for a haplotype) and 0 for other situations. The expectation and maximisation steps are repeated until haplotype frequencies do not significantly change between iterations.

Haplotype distribution within populations and extent of sharing between populations was determined. Measures of diversity at a gene were estimated by considering each haplotype as a separate allele. Phylogenetic analysis was performed to assess evolutionary relationships between haplotypes. Haplotypes generated for each population were collated, and a distance matrix was created for all haplotypes using the GENDIST program. From this matrix a Neighbor-Joining tree was created using NEIGHBOR. GENDIST and NEIGHBOR form part of the Phylip version 3.6 software package(Felsenstein, 2004).

Haplotype diversity was calculated by considering each ML haplotype as a separate allele for a given gene.

$$\hat{H} = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right)$$

where n represents the total chromosomes, k is the number of different haplotypes and p_i is the sample frequency of haplotype i (Nei, 1987).

3.2.4.6 Tests of Selective Neutrality

Mismatch distributions are distributions of the number of differences between pairs of haplotypes and were generated in Arlequin v2.000. Tajima's D contrasts gene diversity calculated under the infinite-sites model to gene diversity calculated under the infinite-alleles model and in Arlequin is calculated from

$$D = \frac{\theta_\pi - \theta_S}{\sqrt{\mathrm{Var}\,(\theta_\pi - \theta_S)}}$$

where $\theta_\pi = \pi$ and $\theta_S = S / \sum_{i=0}^{n-1} (1/i)$ and S is the number of polymorphic loci and i is the number of genes.

Fu's Fs statistic, based on the infinite-sites model, contrasts the probability that a randomly selected neutral population will have more polymorphic sites than the observed population. In Arlequin this is calculated from;

$$F_S = \ln\left(\frac{S'}{1-S'}\right)$$

where S' is the probability that the number of alleles in a neutral population is greater than in the observed population, given the number of observed pairwise differences.

3.2.4.7 F_{ST} Distribution

Population F_{ST}s were generated across all five populations for each of 395 SNPs using the GENETIX v.4.04 program (Belkhir).

3.3 Results

3.3.1 Gene Diversity

Gene diversities were calculated for each population at each gene, and averaged over all 17 genomic regions, using SNP subset 1 (a total of 243 SNPs), as shown in table 3.3.1. Diversity was highest in Antioquia (0.2248) and lowest in Ticuna (0.1635).These results show Beni to be only the third most diverse population behind Spain and Antioquia, which is not consistent with previous studies if Beni represents an average African population.

Average diversities for each gene ranged from 0.1260 (*SLC6A12*) to 0.2919 (*SLC2A3*). Interestingly, two of the most diverse genes, *SLC2A3* and *PA12*, have the smallest deviations about their means demonstrating consistently high diversity in each population. When the population values for *SLC2A3* and *PA12* were compared to the mean population values over all other genes using a Mann-Whitney comparison of means, significance was achieved for each (U=0, p=0.01 for both genes).

3.3.2 Population Structure

Pairwise F_{ST}s between populations were calculated for each of the genomic regions using SNP subset 1 (i.e. 243 SNPs). Mean values for these seventeen estimates are tabulated in table 3.3.2, and similarly average p values are shown in table 3.3.3. An F_{ST} value of 0 signifies no differentiation between populations whereas an F_{ST} of 1 indicates complete differentiation. Antioquia is closest to Spain, as only 1.5% of the total genetic variation across these populations is accounted for by population divergence, and most distant from Ticuna (F_{ST}=13.3). In fact, the most similar populations overall are Antioquia and Spain and are the only two populations not to be significantly different from one another (table 3.3.2). The greatest genetic divergence was found between Ticuna and Beni, with an F_{ST} of 20.8, closely followed by Ticuna and Spain (F_{ST}=19.4). These values can be compared to the mean marker F_{ST} across all five populations of 12.4.

Table 3.3.1. Average gene diversity over all marker loci, π, for 17 genomic regions.

	Beni (36)	Antioquia (63)	Spanish (94)	Chip (43)	Ticuna (48)	Mean	SD
ACVRIIB	0.1540	0.1873	0.1629	0.1128	0.1420	0.1518	0.0274
TGFBI	0.2154	0.2201	0.2096	0.1647	0.1932	0.2006	0.0225
DDR1	0.2019	0.1674	0.1336	0.1812	0.0956	0.1559	0.0419
GTF2H4	0.1955	0.3009	0.3062	0.1940	0.1432	0.2280	0.0722
COLIIA2	0.1085	0.1853	0.1917	0.1423	0.1312	0.1518	0.0357
LAMB1	0.1917	0.3178	0.3167	0.3107	0.2873	0.2848	0.0535
WASL	0.2143	0.1932	0.1513	0.1758	0.1816	0.1832	0.0231
SLC6A12	0.1788	0.1408	0.1330	0.1045	0.0730	0.1260	0.0398
KCNA1	0.1743	0.2316	0.2430	0.2458	0.2621	0.2314	0.0337
SLC2A3 *	0.2832	0.3162	0.2907	0.2774	0.2921	0.2919	0.0148
PCI	0.2411	0.2089	0.2491	0.1078	0.0209	0.1656	0.0985
PRKCBI	0.2462	0.1992	0.2160	0.1528	0.1045	0.1837	0.0557
NF1	0.1940	0.1654	0.1370	0.1211	0.1258	0.1487	0.0306
SCYA2	0.1923	0.1999	0.2139	0.0938	0.0592	0.1518	0.0703
PA12 *	0.2850	0.2704	0.2363	0.2662	0.2666	0.2649	0.0177
IL17R	0.2011	0.2661	0.2417	0.2694	0.1534	0.2263	0.0491
HCF2	0.1890	0.2510	0.2349	0.2285	0.2476	0.2302	0.0248
MEAN	0.2039	0.2248	0.2157	0.1852	0.1635	0.1986	0.0512
SD	0.0433	0.0540	0.0585	0.0694	0.0835		

SD: standard deviation. *Mean population averages for these genes were significantly different when compared to the average of all others, using a Mann-Whitney test (U=0, p=0.01). Chip: Chipewayan.

Table 3.3.2. Pairwise population F_{ST}s averaged over 17 genomic regions.

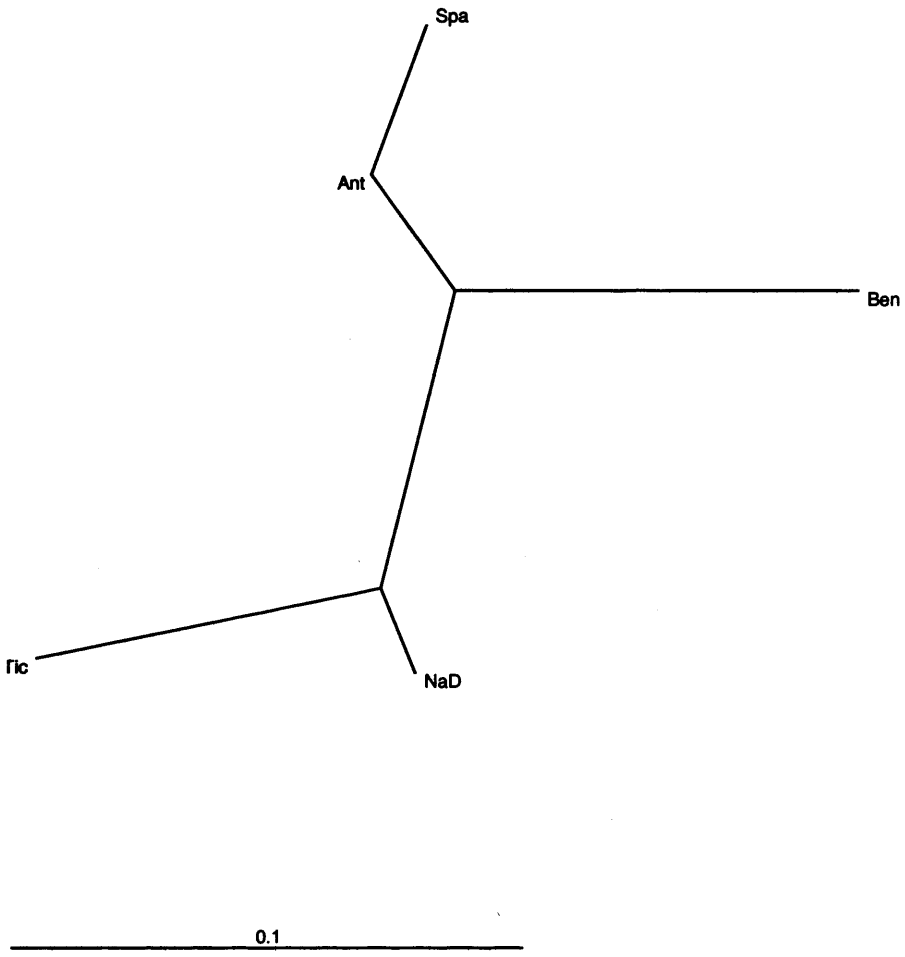
		F_{ST}s			
Pop (n)		**1**	**2**	**3**	**4**
Beni (36)	**1**				
Antioquia (63)	**2**	0.097			
Chipewayan (43)	**3**	0.155	0.083		
Spanish (94)	**4**	0.130	0.015	0.146	
Ticuna (48)	**5**	0.208	0.133	0.087	0.194

Table 3.3.3. Pairwise population F_{ST} p values averaged over 17 genomic regions.

		F_{ST} p values			
Pop (n)		**1**	**2**	**3**	**4**
Beni (36)	**1**				
Antioquia (63)	**2**	0.013			
Chipewayan (43)	**3**	0.012	0.012		
Spanish (94)	**4**	0.002	0.176	0.024	
Ticuna (48)	**5**	0.014	0.023	0.039	0.014

Pairwise F_{ST}s can also represent genetic distance between populations and a Neighbor-Joining tree relating these populations based on F_{ST} distance is shown in figure 3.3.1. The close genetic relatedness of Antioquia and Spain is illustrated, whereas Ticuna appears to be genetically distant from Spain and Beni. Interestingly, Antioquia seems to link Spain with the other populations.

Figure 3.3.1. Neighbour-joining tree based on pairwise F_{ST} genetic distance.



3.3.3 Linkage Disequilibrium

Pairwise LD was assessed between the most frequent minor allele frequency SNPs at each distance bin. Only one SNP was used at each distance to avoid assessing LD in overlapping fragments, which would make measurements not independent. Lewontin's D' was used as a measure of LD and statistical significance was assessed with Fisher's exact test (FET). Tables 3.3.4 and 3.3.5 summarise these findings, and are illustrated in figures 3.3.2 to 3.3.7 and figure 3.3.8. The fifteen unlinked measurements were made by comparing the marker at 10 kb from the core SNP between genes in the Spanish population. The Spanish population was chosen as it represents an old, large, outbred population with high minor allele frequencies where LD is not expected to be strong.

Table 3.3.4. Mean pairwise LD measurements using D', averaged across all 17 loci for each population.

Distance		**5**	**10**	**20**	**40**	**80**	**160**	**Unlinked**
No. Measurements		**40**	**26**	**25**	**21**	**10**	**9**	**15**
	Population							
	Beni	0.855	0.883	0.824	0.608	0.733	0.345	0.318
D'	Antioquia	0.896	0.859	0.779	0.760	0.483	0.207	0.318
	Chip	0.948	0.875	0.780	0.877	0.643	0.423	0.318
	Spanish	0.879	0.880	0.738	0.665	0.464	0.511	0.318
	Ticuna	0.977	0.931	0.915	0.775	0.936	0.547	0.318

Chip: Chipewayan.

Table 3.3.5. The percentage of pairwise LD measurements with FET p value <0.05, averaged across all 17 loci for each population.

Distance		**5**	**10**	**20**	**40**	**80**	**160**	**Unlinked**
No. Measurements		**40**	**26**	**25**	**21**	**10**	**9**	**15**
	Population							
% measurements	Beni	0.600	0.692	0.478	0.238	0.200	0.000	0.000
In LD	Antioquia	0.850	0.885	0.783	0.619	0.200	0.000	0.000
(p<0.05)	Chip	0.875	0.846	0.696	0.667	0.300	0.000	0.000
	Spanish	0.875	0.885	0.739	0.476	0.100	0.000	0.000
	Ticuna	0.800	0.769	0.696	0.571	0.400	0.222	0.000

Chip: Chipewayan.

Figure 3.3.2. Pairwise LD in Beni.

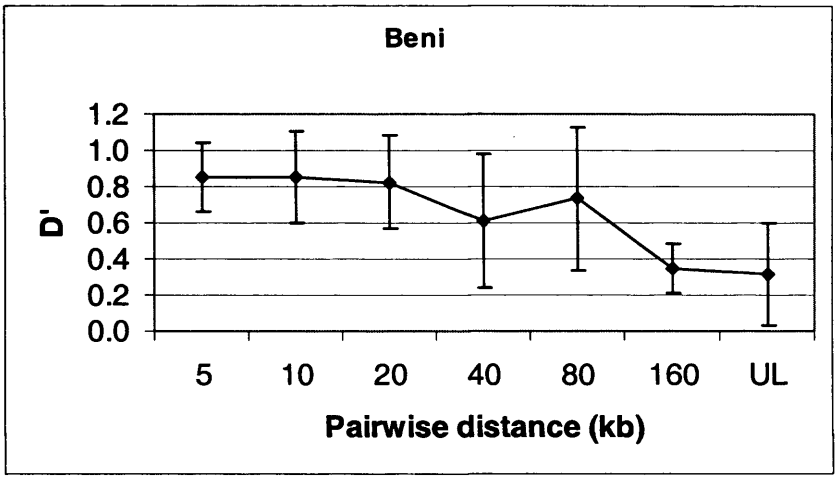


Figure 3.3.3. Pairwise LD in the Spanish population.

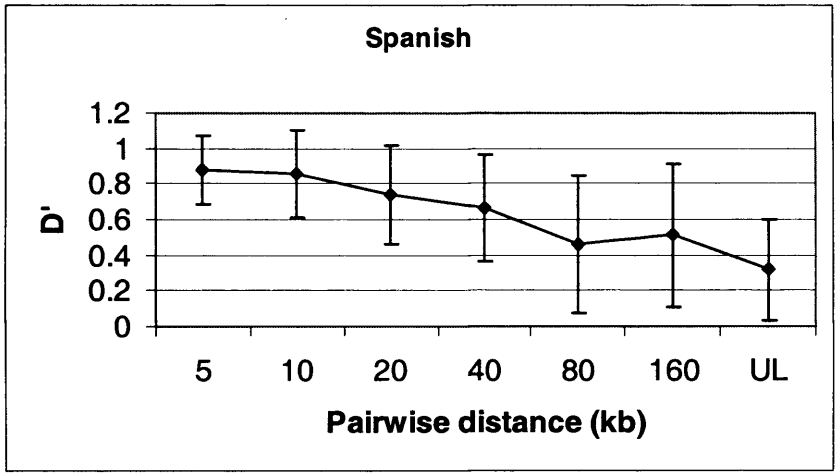


Figure 3.3.4. Pairwise LD in Antioquia

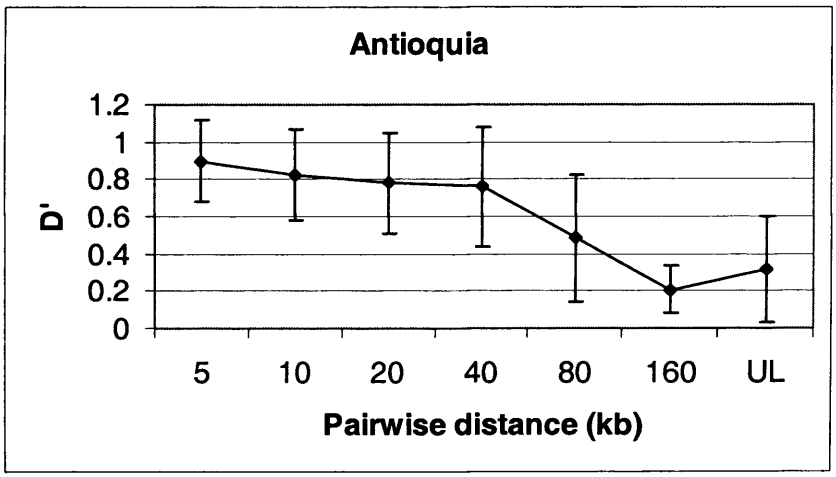


Figure 3.3.5. Pairwise LD in Chipewayan

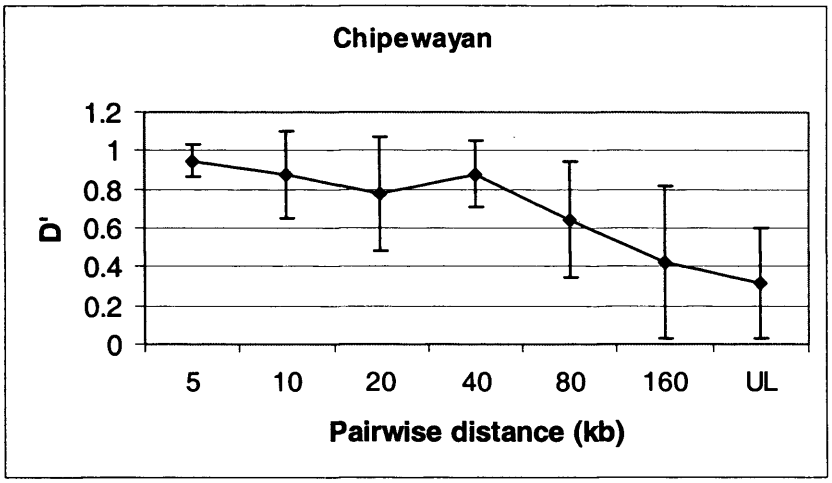
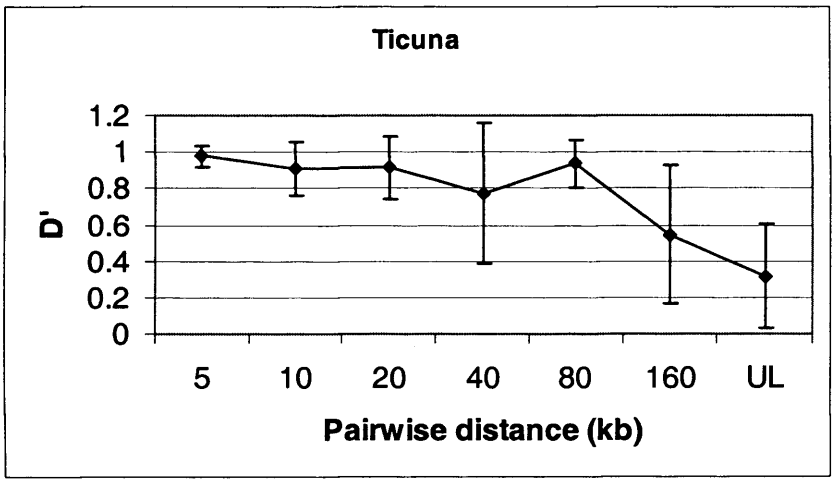
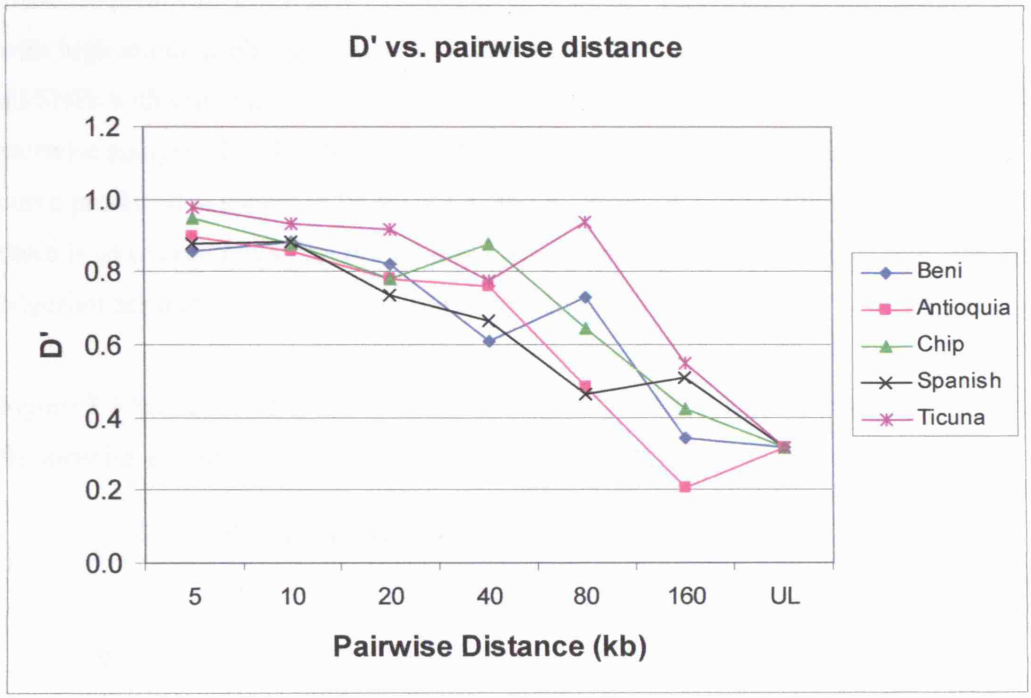


Figure 3.3.6. Pairwise LD in the Ticuna.



Figures 3.3.2 to 3.3.6. Pairwise LD between SNPs at 5, 10, 20, 40, 80 and 160kbs apart. Vertical bars represent standard deviations. D': Lewontin's D prime; UL: unlinked.

Figure 3.3.7. Graph showing D' values at various distances from a core SNP, averaged over 17 autosomal genomic regions, for 5 populations.



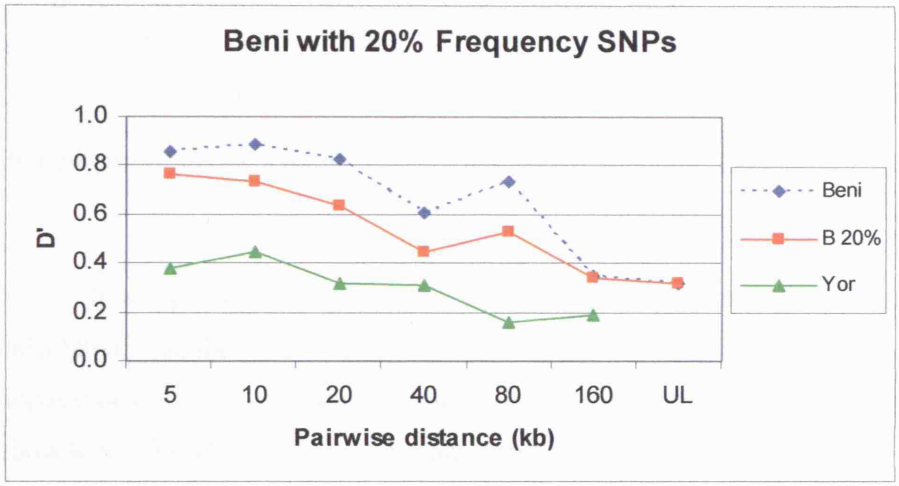
UL: unlinked.

Figures 3.3.2 to 3.3.6 illustrate the trends in D' for each population. These graphs show a lot of variation about means at each distance which may be due to small sample sizes or may represent true patterns of genomic LD distribution. Observations can also be made about population LD, as summarised in figure 3.3.7. The Ticuna seem to maintain strong LD over the greatest distance with a mean D' of 0.936 at 80kb and > 0.5 at 160kb. The Beni and Spanish populations gradually lose their LD from about 20kb, although in both cases uncharacteristic jumps exist at a greater distance. Antioquia's situation may be intermediate between that of the old, outbred populations and the Amerindians. Their D' remains relatively high until 40kb, where it has a value of 0.760, but then drops steadily until 160kb where Antioquia actually shows the lowest LD (and lower than the unlinked sample).

The Beni (from Nigeria) appear to have produced notably different LD estimates than for a Nigerian population in Reich's original work. For example, Reich's work showed the Nigerian population to have a D' of less than 0.4 at 5kb

from the core SNP, whereas I have shown the Beni to have a D' of 0.85 at 5kb. As well as using a core SNP with a minor allele frequency of 35% (in a multi-ethnic panel), the SNPs Reich and colleagues used at the various distance bins had a minor allele frequency of 15/88 in a Utah population. To see what effect using only SNPs with high minor allele frequencies would have on the LD curve for Beni, I removed all SNPs with minor allele frequencies less than 20% (in Beni), and repeated the pairwise analysis. Firstly, this drastically reduced the number of measurements. The curve produced is shown in figure 3.3.8, and it can be seen from this that although there is an overall reduction in LD, it is still considerably higher than obtained for the Nigerian population in Reich's initial analysis.

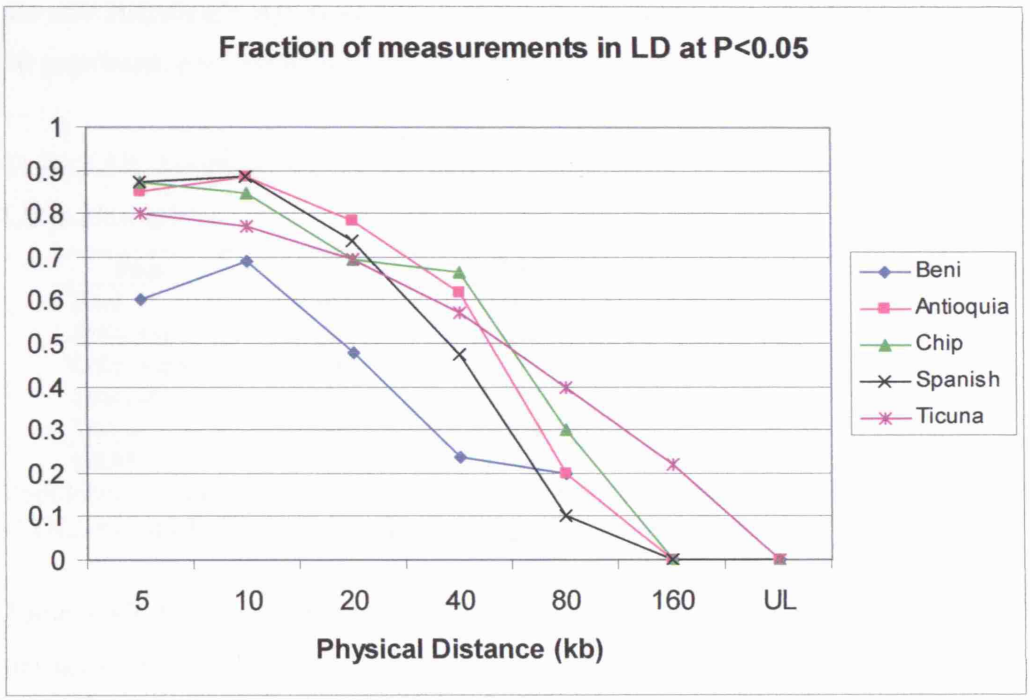
Figure 3.3.8.The effect on D' in Beni of using only SNPs with minor allele frequencies greater than 20%.



The dotted line represents the original curve that used all SNPs. Triangles are an approximation of Nigerian Yoruba population from Reich *et al* (2001). Yor: Yoruban; B 20%: Beni using only ≥ 20%MAF SNPs; UL: unlinked.

Figure 3.3.9 plots the proportion of measurements with FET p values less than 0.05 versus physical distance, often used in other studies as a rough measure of LD. The trends previously suggested with the D' figures are more clearly illustrated here. For example, Beni shows consistently low levels of significant association. Ticuna and Chipewayan maintain relatively strong LD, most extensive in Ticuna. Antioquia has quite strong LD at short distances, until levels drop off at approximately 40kb. Although Antioquia demonstrates consistently more LD than the Spanish the difference is moderate.

Figure 3.3.9. Graph showing percentage of pairwise measurements in significant LD (FET p<0.05) at various distances from a core SNP, for the 5 populations.



Plotted points are the average over 17 autosomal genomic regions.

It is apparent that a discrepancy exists between the molecular diversities calculated from the fuller dataset (table 3.3.1) and the observed LD distribution across populations; for non-admixed populations we would expect diverse populations to show low LD and homogenous populations to show increased LD. However, expected differences in LD levels between the populations are not seen. For example, Ticuna were shown to be the least diverse, non-admixed population but were not shown to have substantially more LD than Spain (the most diverse, non-admixed population). In fact Spain has more LD than Beni, even though Beni was less diverse according to table 3.3.1. As the markers used for the LD analysis represent a subset of those used for molecular diversity calculations (100 SNPs), the discrepancy between LD patterns and π may be due to the properties of the LD markers. To test this I re-calculated nucleotide diversity and F_{ST}s for the markers used in LD analysis only. These results are presented in tables 3.3.6, 3.3.7 and 3.3.8. If the observed homogeneity in LD across populations is due to a bias in marker selection for high heterozygosity SNPs, then we would expect higher π values and smaller genetic

distances. Table 3.3.6 shows that substantially higher πs do exist. Although no obvious increases in F_{ST} values are demonstrated, notably only 2 out of 10 distances are now statistically significant, compared with the larger datasets in which 9 out of 10 population pairs are statistically differentiated.

Table 3.3.6. Summary of gene diversities averaged over 17 genomic regions using the LD marker subset.

Pop	π	SD
Beni	0.368	0.065
Antioquia	0.445	0.048
Chipewayan	0.376	0.071
Spanish	0.420	0.051
Ticuna	0.348	0.109
MEAN	0.391	0.069

Population sample sizes are equal for each region and vary from 19 (*DDR1*) to 32 (*COLIIA2* and *PCI*). π: molecular diversity; SD: standard deviation.

Table 3.3.7. Pairwise population F_{ST}s averaged over 17 genomic regions for the LD marker subset.

		F_{ST}			
		1	2	3	4
Beni	1				
Antioquia	2	0.123			
Chipewayan	3	0.187	0.094		
Spanish	4	0.146	0.009	0.167	
Ticuna	5	0.210	0.135	0.083	0.201

Population sample sizes are equal for each region and vary from 19 (*DDR1*) to 32 (*COLIIA2* and *PCI*).

Table 3.3.8. Pairwise population F_{ST} p values averaged over 17 genomic regions for the LD marker subset.

		F_{ST} p values			
		1	2	3	4
Beni	1				
Antioquia	2	0.089			
Chipewayan	3	0.014	0.107		
Spanish	4	0.041	0.409	0.120	
Ticuna	5	0.063	0.094	0.100	0.091

Population sample sizes are equal for each region and vary from 19 (*DDR1*) to 32 (*COLIIA2* and *PCI*).

3.3.3.1 Variation in LD Across Gene Regions

To investigate genomic variability of LD further, I averaged FET p values at each distance bin over the five populations (table 3.3.9), for each gene. From this analysis it is obvious that some of these genes show very low amounts of LD such as *LAMB1*, *KCNA1*, *PCI* and *IL17R*. Other genes show extensive LD including *DDR1*, *WASL*, *NF1* and *HCF2*.

Table 3.3.9. Fraction of pairwise measurements in significant LD ($p<0.05$) at each distance bin for each gene, averaged over all 5 populations.

	Distance between markers (kb)					
	5	10	20	40	80	160
ACVR2B	0.90	0.80				0.00
TGFB	0.93	1.00	0.80	0.80	0.00	
DDR1*	0.93	1.00	1.00	0.60	0.40	
GTF2H4	0.60	0.60	0.40	0.50	0.20	
COL11A2		0.80	0.60	1.00		
LAMB1	0.50	0.50	0.40	0.00		0.00
WASL*	0.93	1.00	1.00	1.00		
SLC6A12	0.80	0.70	0.20			0.20
KCNA1	0.40	0.40		0.00	0.20	0.00
SLC2A3	0.90	0.70	0.70	0.60		0.00
PCI	0.20			0.00		0.00
PRKCBI	1.00	1.00	0.80	0.40	0.00	0.00
SCYA2	0.90	1.00	0.60	0.40	0.20	
NF1*	1.00	1.00	0.90	0.90	1.00	
PA12	1.00	0.90	0.90	0.40	0.20	
IL17R	0.30	0.20	0.00			0.00
HCF2*	1.00	1.00	1.00	1.00		0.20

* represents regions of high LD selected for haplotype analysis.

3.3.4 Haplotype Analysis

Maximum likelihood haplotypes were generated from the four genomic regions that showed the highest degree of linkage disequilibrium (*DDR1*, *WASL*, *NF1* and *HCF2*). Markers were selected from subset 1 on the basis of heterozygosity and level of genotyping success (see methods). Haplotypes consisted of thirteen markers for *DDR1*, six for *WASL*, nine for *NF1* and twelve for *HCF2* and sample sizes varied depending on gene and population (ranging from 20 Beni for *DDR1* to 72 Spanish for *WASL*).

3.3.4.1 Frequency Distribution

From the frequency distribution graphs (figures 3.3.11 to 3.3.15, 3.3.17 to 3.3.21, 3.3.23 to 3.3.27 and 3.3.29 to 3.3.33) it is clear that the Beni consistently have a more even distribution of haplotypes. This results in a consistently high haplotypic diversity for the Beni with a mean of 0.790 (table 3.3.10). Ticuna, on the other hand, tend to have distributions dominated by one or two haplotypes and have the lowest mean diversity for the populations examined at 0.559.

Table 3.3.10. Haplotype diversity at four gene regions for each population.

	Haplotype Diversity				
	DDR1	**WASL**	**NF1**	**HCF2**	**MEAN**
Beni	0.835	0.702	0.753	0.869	0.790
Antioquia	0.819	0.659	0.525	0.828	0.708
Spanish	0.798	0.661	0.326	0.826	0.653
Chipewayan	0.745	0.685	0.494	0.814	0.685
Ticuna	0.395	0.673	0.460	0.706	0.559
MEAN	0.718	0.676	0.512	0.809	0.679

The underlined value shows an uncharacteristically low diversity in Spain.

The Chipewayan and Antioquia show strong increases in haplotype diversity compared to the Ticuna, with mean values of 0.685 and 0.708 respectively. The mean Spanish haplotype diversity is lower than expected due to an unusually low diversity value at the *NF1* gene, while values are high at all other genes.

3.3.4.2 Haplotype Sharing Between Populations

The pattern of haplotype sharing between Spain, Antioquia, Chipewayan and Ticuna is consistent with their historical relationships. If looking at haplotypes that are very common in Spain and rare in Ticuna the following trend is repeated several times: the 'Spanish' haplotype is second most frequent in the Antioquians, followed by the Chipewayan showing much less of the haplotype, and finally the Ticuna with trace amounts of the Spanish haplotype. This is seen with the following haplotypes: DR1, WASL2, NF4 and HCF5 (figures 3.3.10, 3.3.16, 3.3.22 and 3.3.28 respectively). What's more, when looking at haplotypes that are common in Ticuna and rare in Spain the reverse trend exists; for example, haplotypes DR8, NF2 and HCF9 (figures 3.3.10, 3.3.22 and 3.3.28 respectively). Interestingly, these results show that whereas the Ticuna appear relatively separate from the Spanish, there is a considerable number of Spanish haplotypes in the Chipewayan. Additionally, some haplotypes exist at low

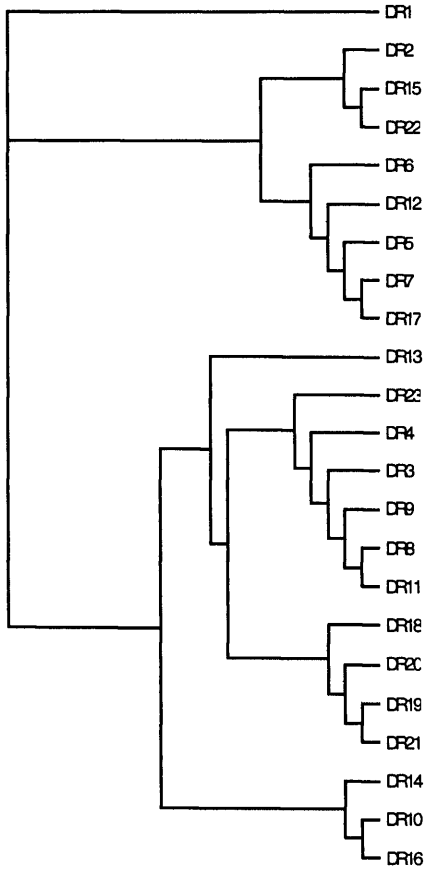
frequencies in both the Beni and Antioquia, likely representing the small African contribution to present day Antioquia.

3.3.4.3 Haplotype Phylogeny

Phylogenetic analysis has grouped the haplotypes together based on their sequence similarity into clades or 'haplogroups'. Some haplogroups are present only in Antioquia and Spain, such as the bottom haplogroup of *DDR1* (figure 3.3.10), augmenting previous findings regarding the strong genetic similarity between the two populations. Others, like the top haplogroup in *HCF2* (containing HCF8, 22, 15 and 17 – figure 3.3.28), are only absent in Chipewayan and Ticuna. Additionally, only Chipewayan and Ticuna possess the group containing haplotypes HCF24, 11, 18 and 19, apart from HCF11 which is present at low frequencies in Antioquia (0.03). This shows how the *HCF2* and *DDR1* haplogroups are able to distinguish between three major, continental groups: America, Europe and Africa. What's more, one *DDR1* haplogroup, containing haplotypes DR18 – DR21 figure (3.3.10), is present only in the Chipewayan population.

An intriguing observation can be made from the *NF1* haplotype tree (figure 3.3.22). In this tree there is one haplogroup, containing only haplotype NF9, that is present at a frequency of 0.208 in the Ticuna, 0.06 in the Chipewayan and 0.049 in the Antioquian population, but absent in both the Spanish and the Beni. Furthermore, there exists another haplogroup - consisting of haplotypes NF4 and NF11 - that is very common in both Spain and Antioquia, at moderate levels in the Chipewayan but rare in the Beni. As the remaining haplogroup for the *NF1* gene contains haplotype NF2, common to all populations and likely resulting from the Upper Palaeolithic expansion, these findings may suggest a non-African contribution to genetic variation seen here.

Figure 3.3.10. Neighbour-joining tree of maximum likelihood LD haplotypes for *DDR1*.



		Beni	Ant	Span	Chip	Tic	Mean
DR1	xxoooxoooxoox	0.325	0.181	0.229	0.074	0.000	0.162
DR2	xoooooxoooxoxo	0.175	0.056	0.119	0.132	0.000	0.096
DR15	oxoooxoooxoxo	0.000	0.014	0.000	0.000	0.000	0.003
DR22	xxoooxoooxoxo	0.000	0.000	0.017	0.000	0.000	0.003
DR6	xoxooxoooxoox	0.050	0.069	0.008	0.000	0.000	0.025
DR12	ooxooxoooxoox	0.000	0.014	0.008	0.000	0.000	0.004
DR5	xoxxoxoooxoox	0.075	0.250	0.322	0.103	0.153	0.181
DR7	ooxxoxoooxoox	0.050	0.000	0.000	0.000	0.000	0.010
DR17	xoxxxxoooxoox	0.000	0.000	0.000	0.000	0.014	0.003
DR13	oxoooxoooxoox	0.000	0.014	0.017	0.000	0.000	0.006
DR23	xxooooxxoooox	0.000	0.000	0.000	0.000	0.014	0.003
DR4	xxoooxxxxoox	0.125	0.014	0.008	0.000	0.000	0.029
DR3	oxoooxxxxoox	0.150	0.000	0.008	0.000	0.000	0.032
DR9	xoxooxxxxoox	0.025	0.000	0.000	0.000	0.000	0.005
DR8	xxoooxxxxoox	0.025	0.292	0.169	0.441	0.764	0.338
DR11	xxooxoxxxoox	0.000	0.014	0.000	0.191	0.056	0.052
DR18	xxoooxxxooox	0.000	0.000	0.000	0.015	0.000	0.003
DR20	xxooxxxxooox	0.000	0.000	0.000	0.015	0.000	0.003
DR19	xxooxoxxooox	0.000	0.000	0.000	0.015	0.000	0.003
DR21	xxoooxxooox	0.000	0.000	0.000	0.015	0.000	0.003
DR14	oxoooxoooxxox	0.000	0.014	0.000	0.000	0.000	0.003
DR10	xxoooxoooxxox	0.000	0.056	0.093	0.000	0.000	0.030
DR16	xxooxxoooxxox	0.000	0.014	0.000	0.000	0.000	0.003
	Diversity	0.835	0.819	0.798	0.745	0.395	

Haplotype frequencies in each of the five populations are shown, with an estimation of haplotype diversity. Haplotypes with frequencies < 0.005 are considered as 0. Ant: Antioquia; Chip: Chipewayan; Span: Spanish; Tic: Ticuna.

Figures 3.3.11 to 3.3.15. Population frequency distributions for the *DDR1* gene haplotypes.

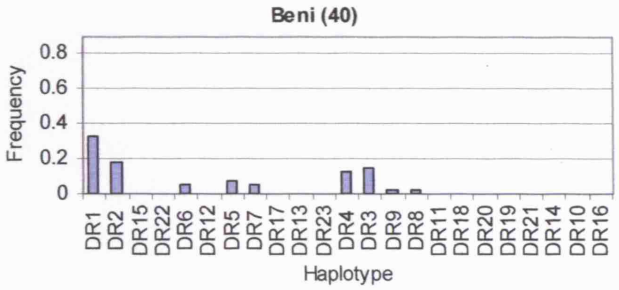


Figure 3.3.11.

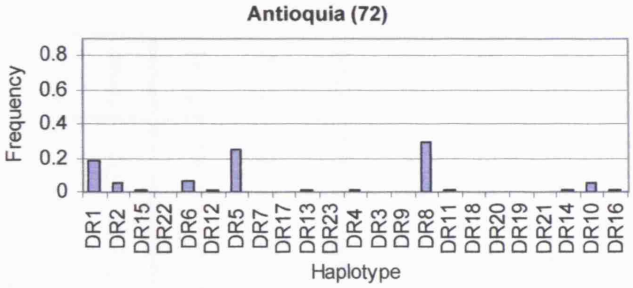


Figure 3.3.12.

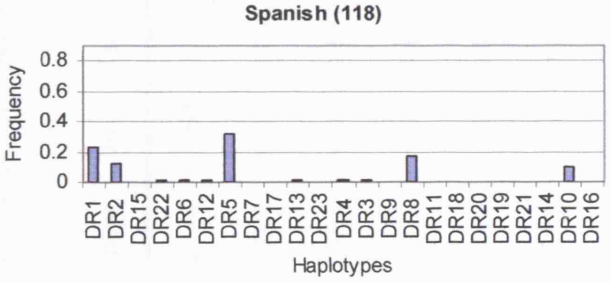


Figure 3.3.13.

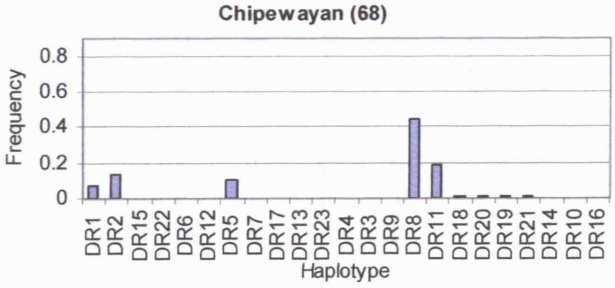


Figure 3.3.14.

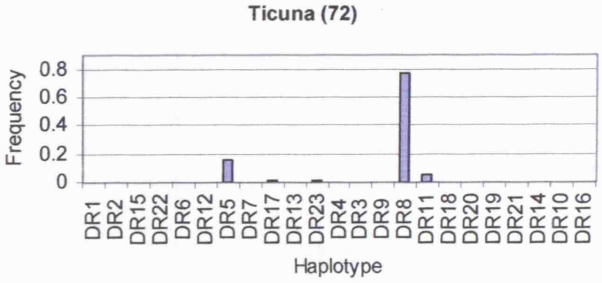
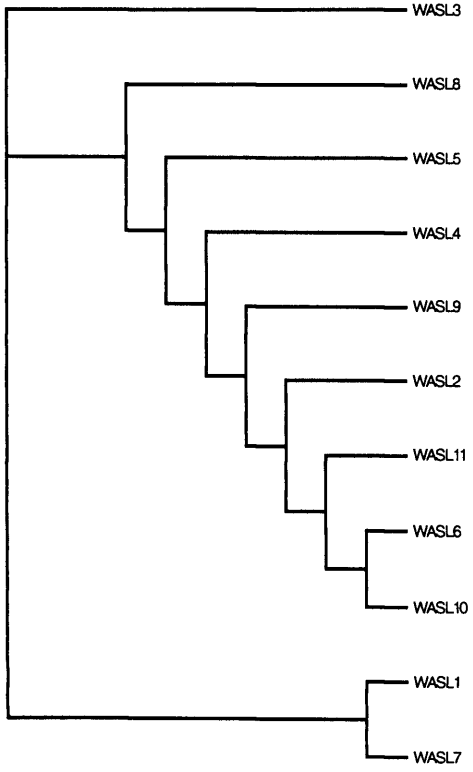


Figure 3.3.15.

Figure 3.3.16. Neighbour-joining tree of maximum likelihood LD haplotypes for *WASL*.



		Beni	Ant	Span	Chip	Tic	Mean
WASL3	oooooo	0.146	0.250	0.294	0.197	0.342	0.246
WASL8	oooxoo	0.000	0.000	0.008	0.000	0.000	0.002
WASL5	oxoooo	0.024	0.000	0.000	0.000	0.000	0.005
WASL4	xxoooo	0.073	0.000	0.000	0.000	0.000	0.015
WASL9	xxoxoo	0.000	0.000	0.008	0.000	0.000	0.002
WASL2	xxoxxo	0.268	0.431	0.462	0.273	0.247	0.336
WASL11	xxooxo	0.000	0.000	0.000	0.000	0.014	0.003
WASL6	xxxxxo	0.024	0.000	0.000	0.000	0.000	0.005
WASL10	oxxoxo	0.000	0.000	0.008	0.000	0.000	0.002
WASL1	ooxoox	0.463	0.319	0.210	0.455	0.397	0.369
WASL7	ooxooo	0.000	0.000	0.008	0.076	0.000	0.017
	Diversity	0.702	0.659	0.661	0.685	0.673	

Haplotype frequencies in each of the five populations are shown, with an estimation of haplotype diversity. Haplotypes with frequencies < 0.005 are considered as 0. Ant: Antioquia; Span: Spanish; Chip: Chipewayan; Tic: Ticuna.

Figures 3.3.17 to 3.3.21. Population frequency distributions for the *WASL* gene haplotypes.

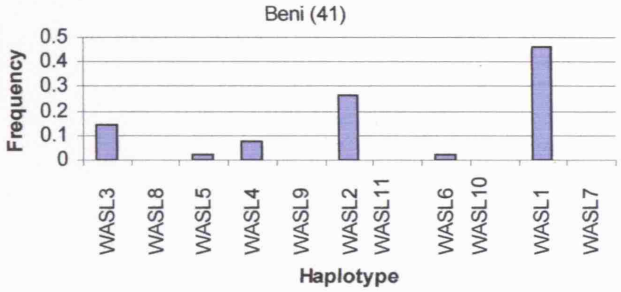


Figure 3.3.17.

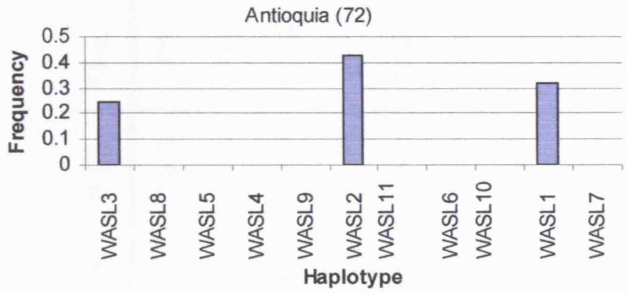


Figure 3.3.18.

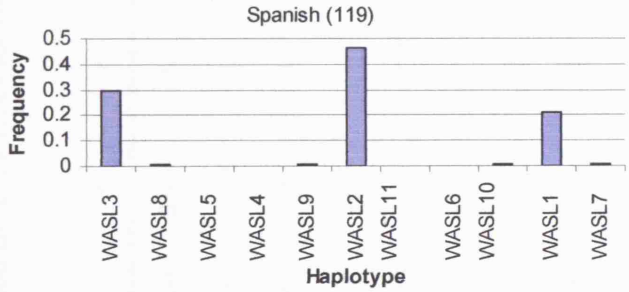


Figure 3.3.19.

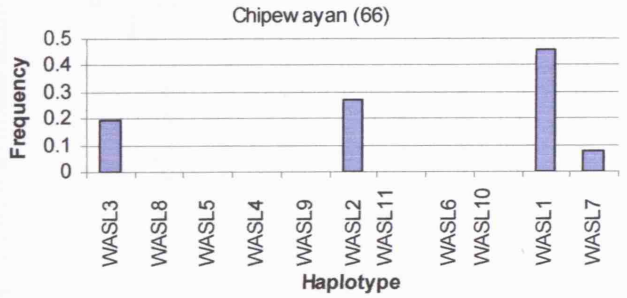


Figure 3.3.20.

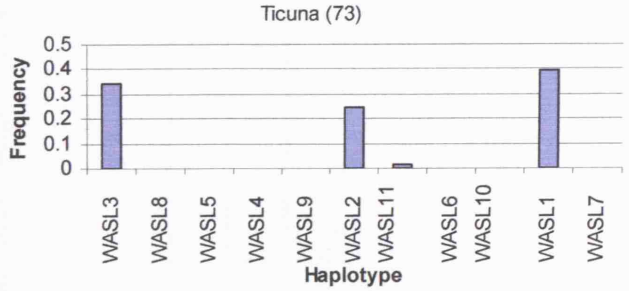
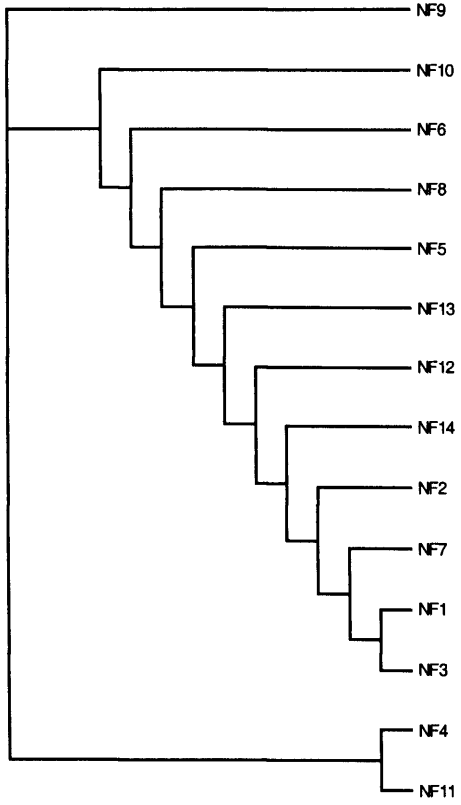


Figure 3.3.21.

Figure 3.3.22. Neighbour-joining tree of maximum likelihood LD haplotypes for *NF1*.



		Beni	Ant	Span	Chip	Ticuna	Mean
NF9	ooxoxoxox	0.000	0.049	0.000	0.060	0.208	0.063
NF10	ooxxooxoo	0.000	0.000	0.000	0.040	0.000	0.008
NF6	oxxoooxxo	0.048	0.000	0.000	0.000	0.000	0.010
NF8	oxxoooxox	0.016	0.000	0.007	0.000	0.000	0.005
NF5	oxxooooox	0.065	0.020	0.000	0.020	0.000	0.021
NF13	xxxooxoxx	0.000	0.000	0.007	0.000	0.000	0.001
NF12	xxxxoxoox	0.000	0.000	0.000	0.020	0.000	0.004
NF14	xoxxoxoxx	0.000	0.000	0.000	0.000	0.021	0.004
NF2	xxxxoxoxx	0.323	0.265	0.164	0.700	0.708	0.432
NF7	xxxxoooox	0.048	0.000	0.007	0.000	0.000	0.011
NF1	xxoxoooxx	0.371	0.029	0.007	0.000	0.000	0.081
NF3	xxxxoooxx	0.065	0.000	0.000	0.000	0.000	0.013
NF4	ooxoxoxoo	0.065	0.637	0.806	0.140	0.063	0.342
NF11	ooxxxoxoo	0.000	0.000	0.000	0.020	0.000	0.004
	Diversity	0.753	0.525	0.326	0.494	0.460	

Haplotype frequencies in each of the five populations are shown, with an estimation of haplotype diversity. Haplotypes with frequencies < 0.005 are considered as 0. Ant: Antioquia; Chip: Chipewayan; Span: Spanish.

Figures 3.3.23 to 3.3.27. Population haplotype frequency distributions for the *NF1* gene.

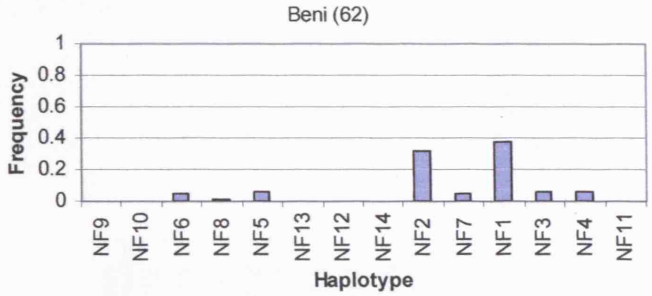


Figure 3.3.23.

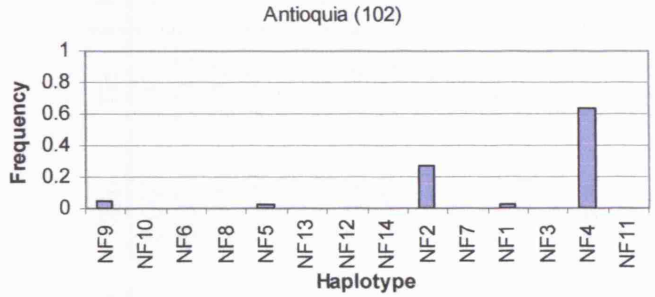


Figure 3.3.24.

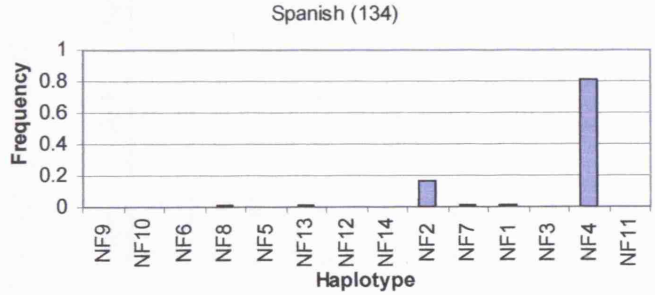


Figure 3.3.25.

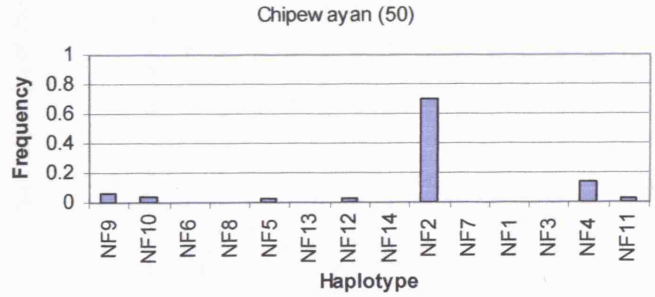


Figure 3.3.26.

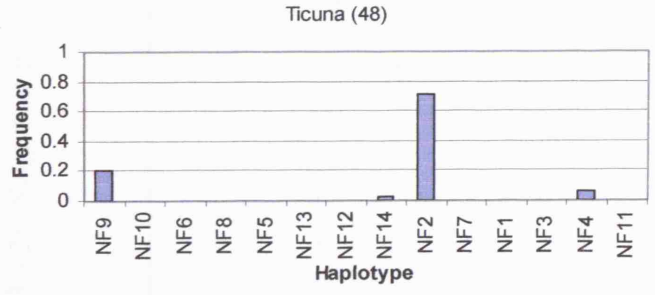
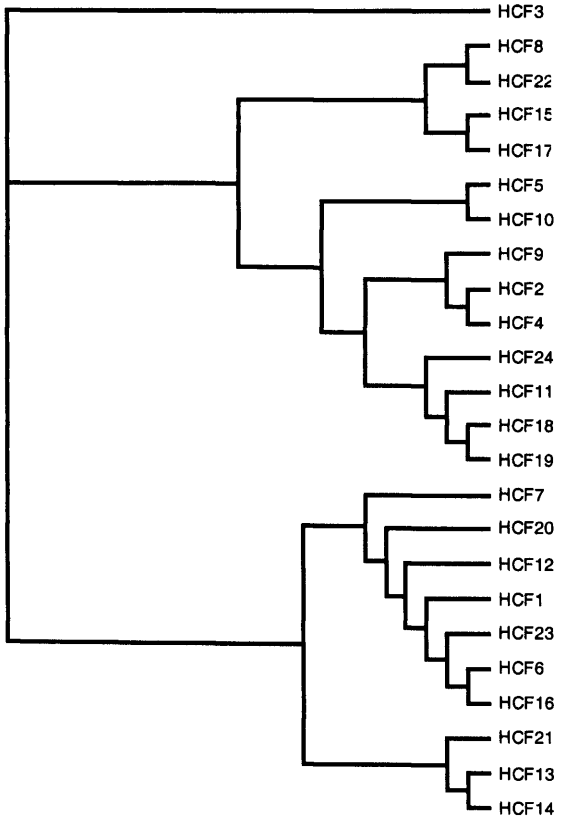


Figure 3.3.27.

Figure 3.3.28. Neighbour-joining tree of maximum likelihood LD haplotypes for *HCF2*.



		Beni	Ant	Span	Chip	Tic	Mean
oxxxxxoxoxox		0.159	0.000	0.000	0.000	0.000	0.032
ooxxxoxxxoox		0.046	0.050	0.076	0.000	0.000	0.034
xoxxxoxxxoox		0.000	0.000	0.017	0.000	0.000	0.003
xxxxxxxoxoox		0.000	0.010	0.000	0.000	0.000	0.002
oxxxxxxoxoox	*	0.000	0.000	0.000	0.000	0.000	0.000
oxxxxxooooox		0.136	0.230	0.277	0.172	0.023	0.168
oxxxxxoooooo		0.000	0.050	0.034	0.000	0.000	0.017
xxxxxoxooooox		0.000	0.090	0.000	0.109	0.432	0.126
xxxxxxooooox		0.159	0.000	0.067	0.063	0.000	0.058
xxxxxxoxoxox		0.136	0.030	0.000	0.000	0.000	0.033
xxxxoxoxoxox		0.000	0.000	0.000	0.000	0.023	0.005
oxxxoxooooox		0.000	0.030	0.000	0.031	0.114	0.035
ooxoxooooox		0.000	0.000	0.000	0.344	0.000	0.069
xooxoxooooox		0.000	0.000	0.000	0.031	0.000	0.006
ooxoxooxoxxx		0.068	0.000	0.000	0.000	0.000	0.014
oooxxoxoxxx		0.000	0.000	0.059	0.000	0.000	0.012
xooxxoxoxxx		0.000	0.020	0.017	0.016	0.000	0.011
oooxoxoxoxxx		0.227	0.260	0.261	0.172	0.318	0.248
xooooooxoxxx		0.000	0.000	0.008	0.000	0.000	0.002
xoooxooxoxxx		0.068	0.210	0.143	0.063	0.091	0.115
xoooxoxxxxx	*	0.000	0.000	0.000	0.000	0.000	0.000
xxxoxooxoxox		0.000	0.000	0.025	0.000	0.000	0.005
oxxoxooxoxox		0.000	0.010	0.008	0.000	0.000	0.004
xxxoxooxoxoo		0.000	0.010	0.008	0.000	0.000	0.004
Diversity		0.869	0.828	0.826	0.814	0.706	

Haplotype frequencies in each of the five populations are shown, with an estimation of haplotype diversity. Haplotypes with frequencies < 0.005 are considered as 0. * haplotypes HCF16 and HCF17 were estimated at frequencies of less than 0.001 in Antioquia. Ant: Antioquia; Chip: Chipewayan; Span: Spanish; Tic: Ticuna.

Figures 3.3.29 to 3.3.33. Population frequency distributions for the *HCF2* gene haplotypes.

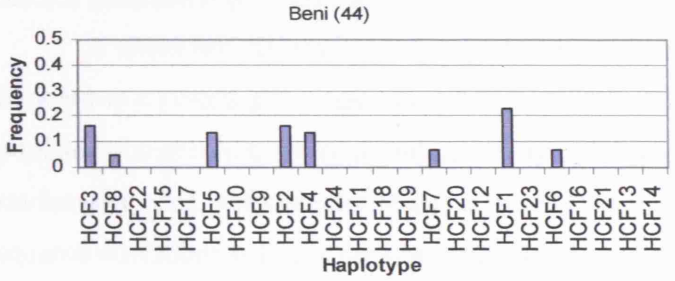


Figure 3.3.29.

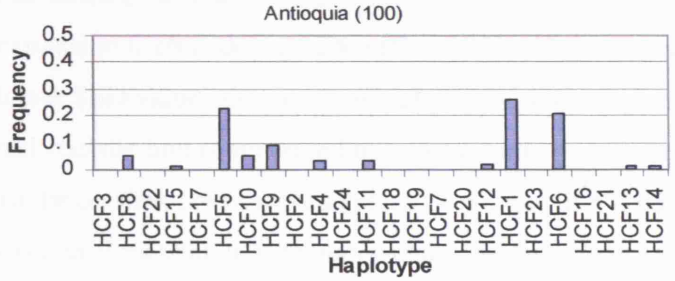


Figure 3.3.30.

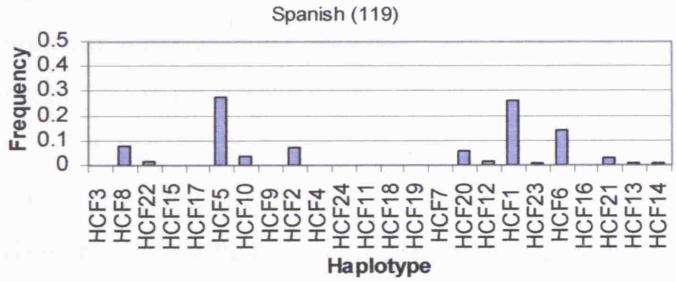


Figure 3.3.31.

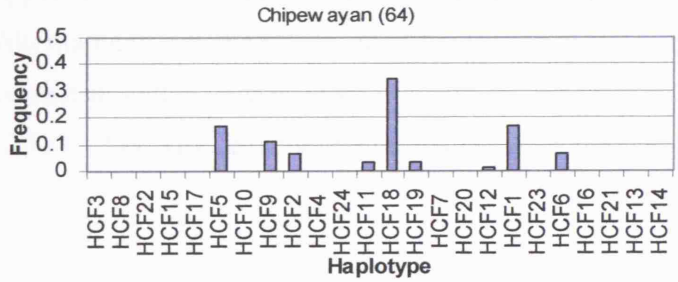


Figure 3.3.32.

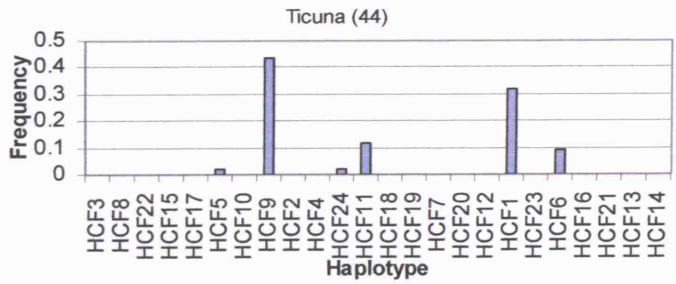


Figure 3.3.33.

3.3.5 Population Expansion

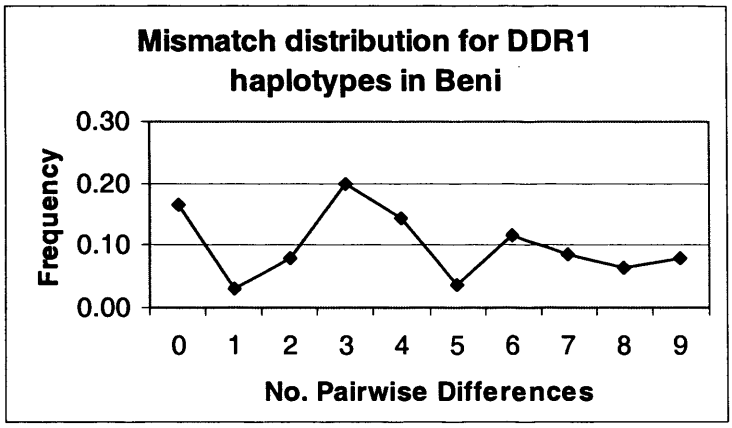
3.3.5.1 Mismatch Distribution

A mismatch distribution refers to the distribution of nucleotide differences between two homologous sequences. A unimodal distribution may signify a historic population expansion, whereas multimodality reflects a state of population stationarity. One measurement of this is the SSD statistic which gives the sums of squared deviations between the observed mismatch distribution and the expected mismatch distribution under a model of population growth, in this case population expansion (Schneider and Excoffier, 1999). Unimodal distributions will generate lower SSD values than multimodal distributions. A p value can be attributed to the test statistic and is generated by determining the number of simulated SSD values that are larger than the observed values. A p value of ≤ 0.05 signifies that 5% or less of the simulated mismatch distributions have more variation than the observed distribution (i.e. 95% or more of the simulated distributions have less variation, and are therefore more unimodal, than the observed), and is taken as a departure from the expansion model (Excoffier and Schneider, 1999; Schneider and Excoffier, 1999).

A significant deviation from a sudden expansion model was demonstrated for most of my populations using the maximum likelihood haplotypes from the four high LD regions. Notable exceptions are for the *DDR1* gene for which all populations appear to follow a sudden expansion model (see figures 3.3.34 to 3.3.38). Alternatively, for *WASL* only the Chipewayan show a mismatch distribution consistent with a sudden expansion (3.3.39), and this trend is repeated with *HCF2* (3.3.40). For *NF1* only the Beni show evidence for a recent expansion (3.3.41).

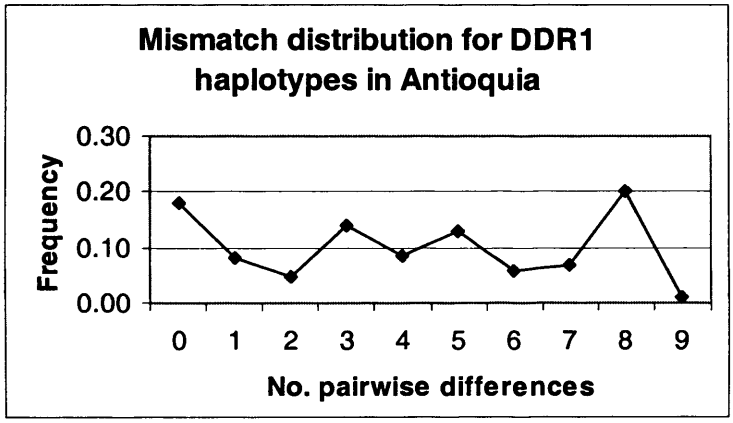
Figures 3.3.34 to 3.3.38. Unimodal mismatch distributions for the *DDR1* haplotypes.

Figure 3.3.34.



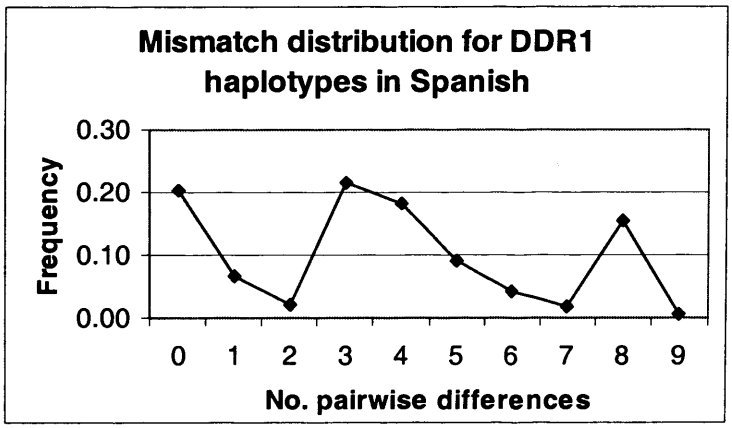
p = 0.210, so no significant departure from a sudden expansion model.

Figure 3.3.35.



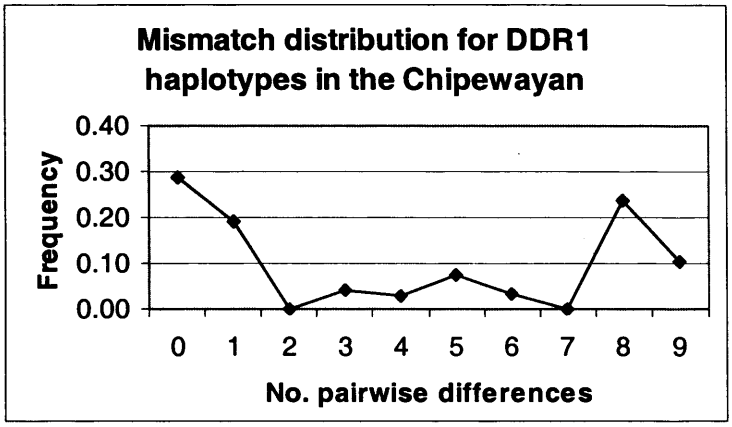
p = 0.17, so no significant departure from a sudden expansion model.

Figure 3.3.36.



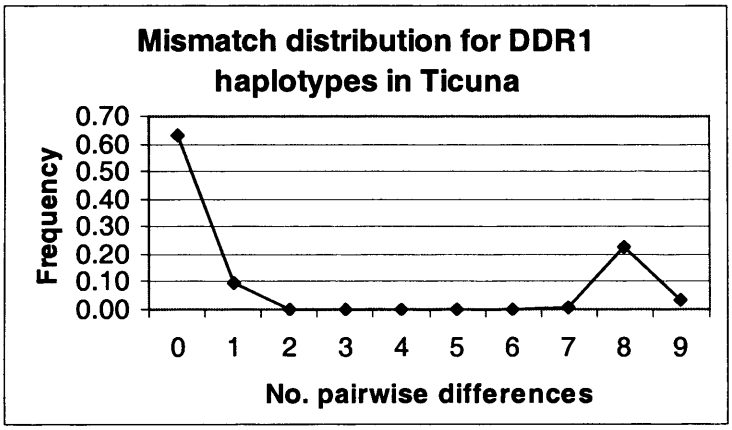
p = 0.07, so no significant departure from a sudden expansion model.

Figure 3.3.37.



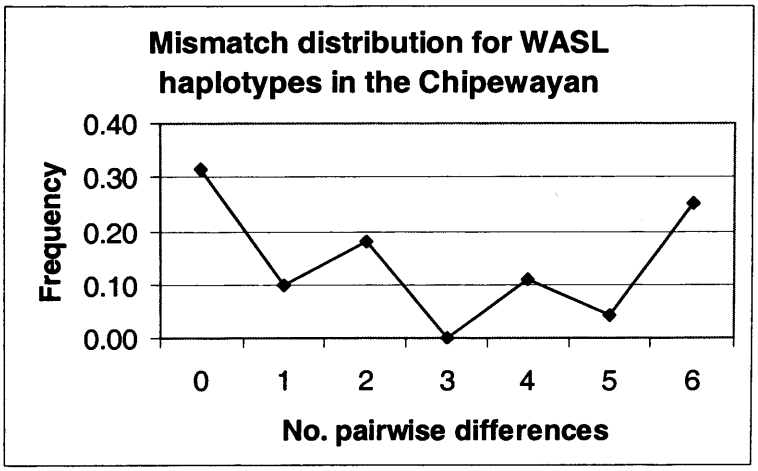
p = 0.11, so no significant departure from a sudden expansion model.

Figure 3.3.38.



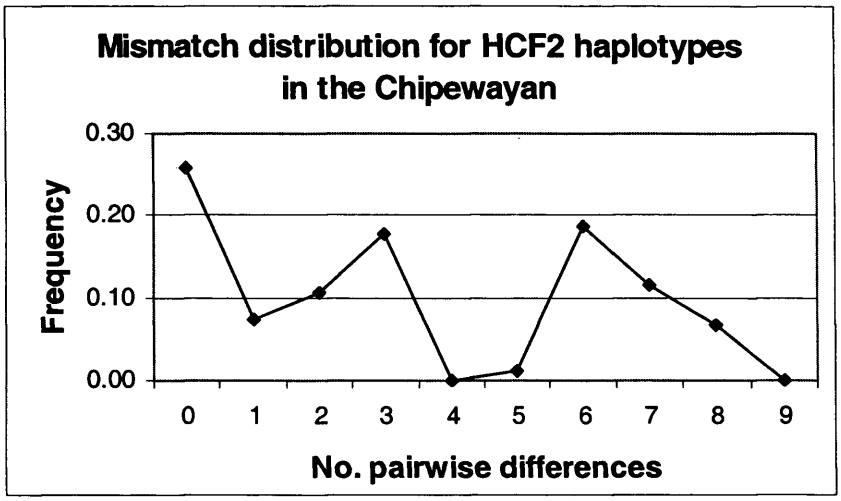
p = 0.08, i.e., no significant departure from a sudden expansion model.

Figure 3.3.39. Mismatch distribution for *WASL* in the Chipewayan.



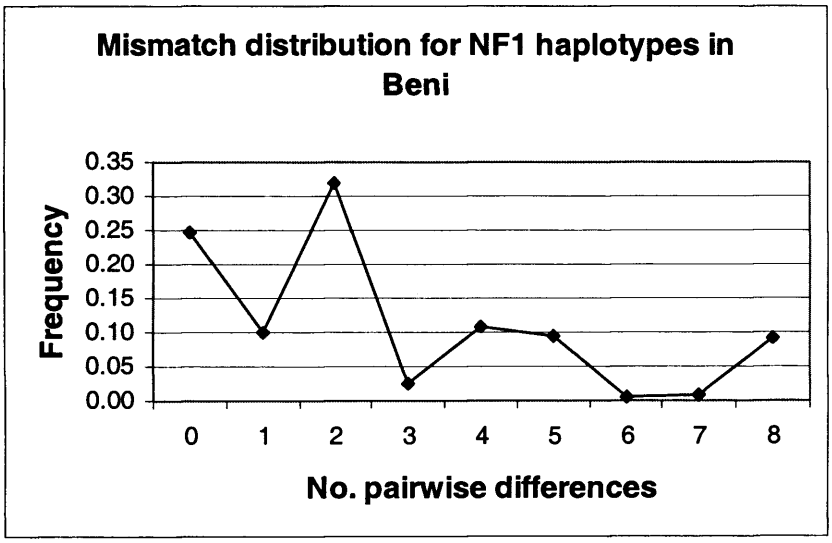
p = 0.07, i.e., no significant departure from a sudden expansion model.

Figure 3.3.40. Mismatch distribution for *HCF2* in the Chipewayan.



p = 0.16, i.e., no significant departure from a sudden expansion model.

Figure 3.3.41. Mismatch distribution for *NF1* in Beni.



p = 0.11, i.e., no significant departure from a sudden expansion model.

3.3.5.2 Tests of Neutrality

Tajima's D and Fu's Fs measures were used to detect for population deviation from neutrality for each of the four genes. Negative values are believed to indicate directional selection, whereas large positive values may suggest balancing selection (Bamshad et al., 2002; Tajima, 1989). No statistically significant deviation from a neutral expectation was observed for any gene in all populations (table 3.3.11). It has been shown, however, that many factors such as mutation rate heterogeneity, degrees

of population expansion and time since expansion can strongly influence Tajima's D values (ArisBrosou and Excoffier, 1996).

Table 3.3.11. Tajima's D and Fu's Fs values based on all markers used to generate ML haplotypes for the high LD genes.

	DDR1				NF1				WASL				HCF2			
	D	p	Fs	p	D	p	Fs	p	D	p	Fs	p	D	p	Fs	p
Beni	1.76	0.93	1.10	0.79	0.89	0.73	0.79	0.75	2.74	0.98	2.14	0.97	1.90	0.94	4.42	0.99
Antioquia	1.48	0.86	-0.09	0.64	2.51	0.97	6.81	1.00	3.13	0.99	7.98	1.00	2.39	0.95	2.47	0.95
Chip	1.56	0.91	3.67	0.96	1.95	0.94	1.95	0.92	2.60	0.97	5.22	1.00	2.55	0.98	5.23	0.99
Spanish	1.59	0.92	1.49	0.83	1.17	0.82	3.58	0.97	3.05	0.98	2.89	0.99	2.30	0.96	4.38	0.98
Ticuna	0.42	0.70	3.19	0.91	1.77	0.93	5.24	1.00	2.39	0.97	5.07	1.00	3.25	1.00	6.98	1.00

p values represent the proportion of simulated values, generated under neutral models of population expansion, less than the observed data. Chip: Chipewayan.

3.3.6 F_{ST} Distribution

Figure 3.3.42. Distribution of F_{ST}s across all five populations for 395 SNPs from seventeen genomic regions.

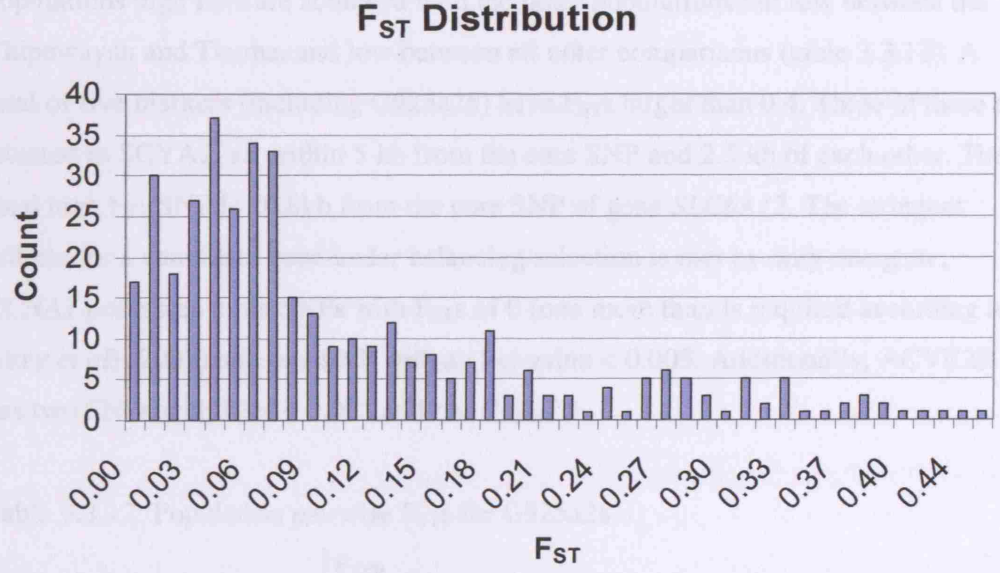


Figure 3.3.42 shows a distribution of F_{ST} values for 395 markers over all five populations. The observed distribution has a mean value of 0.1041 with a standard

deviation of 0.0989. There are four markers with F_{ST} values ≥ 0.4, which represent 1.01% of the distribution; and there are seventeen markers (4.3% of the total distribution) with an F_{ST} value of 0. These proportions of markers that lie in the tail ends of the distribution are less than in a previous study that used 26,530 genome-wide SNPs (Akey et al., 2002). As Weir and Cockerham's unbiased estimate of F_{ST} was used slightly negative values are achievable (Weir and Cockerham, 1984); two were obtained here and considered to be 0. In the study by Akey *et al* (2002) SNPs with F_{ST}s greater than 0.45 gave a significance level of 0.026 based on the empirical distribution and were therefore classified as representing directional selection candidate genes. Balancing selection candidate genes required the inclusion of two SNPs with F_{ST}s of 0 and one with an $F_{ST} \leq 0.005$. These figures are used as a guideline here as both studies have looked at a genome wide selection of SNPs that encompass randomly selected coding, intronic and non-coding regions. Only one SNP from my analysis was shown to have an F_{ST} value larger than 0.45: marker G925a28 which has an F_{ST} of 0.52 and is situated approximately 80kb from the *DDR1* core SNP. Interestingly, without the Ticuna this F_{ST} drops to 0.33 indicating allele frequency differences in the Ticuna are responsible for the inflated F_{ST}. Pairwise F_{ST}s for this marker are more revealing and show that for both Native American populations high F_{ST}s are achieved with the other populations but low between the Chipewayan and Ticuna, and low between all other comparisons (table 3.3.12). A total of five markers (including G925a28) have F_{ST}s larger than 0.4. Three of these are situated in SCYA2, all within 5 kb from the core SNP and 2.5 kb of each other. The final high F_{ST} SNP is 20.8kb from the core SNP of gene *SLC6A12*. The stringent criteria for a candidate gene under balancing selection is met by only one gene; *KCNA1* possesses three SNPs with F_{ST}s of 0 (one more than is required according to Akey *et al*) (2002) and one SNP with an F_{ST} value < 0.005. Additionally, ACVR2B has two SNPs with F_{ST}s < 0.005 and one F_{ST} of 0.

Table 3.3.12. Population pairwise F_{ST}s for G925a28.

		F_{ST}s			
Pop (n)		1	2	3	4
Beni (36)	1				
Antioquia (63)	2	0.093			
Chipewayan (43)	3	0.701	0.429		
Spanish (94)	4	0.000	0.058	0.645	
Ticuna (48)	5	0.851	0.608	0.078	0.773

3.4 Discussion

3.4.1 Antioquia

A major objective of this study was to characterise the autosomal genetic diversity of the Antioquia population isolate from northwest Colombia. This population is a geographic population isolate with a unique demographic history, and as such may have a special role in mapping human disease genes. The genetic structure (due possibly to founding effects) of Antioquia has already enabled positive linkage for several disorders with a simple inheritance pattern including familial forms of Parkinson's Disease and Alzheimer's Disease (Lendon et al., 1997; Pineda-Trujillo et al., 2001). This population may be useful, therefore, in mapping genes for common, complex traits and disorders, which will rely on comprehensive knowledge of its genetic make-up.

Several studies have been carried out to characterise the genetic structure of Antioquia and have revealed traces of past demographic events including an admixture founding event based on Spanish male, Amerind female driven admixture (with a bias towards Spanish males) (Carvajal-Carmona et al., 2000; Sandoval et al., 1993). A comparative analysis of diversity in Antioquia and two other population isolates has been done, including a somewhat limited assessment of linkage disequilibrium, that showed low diversity in the mtDNA but moderate diversity in the Y chromosome (Carvajal-Carmona et al., 2003). However, these studies have concentrated on the uniparentally inherited mtDNA and NRY genetic systems, and therefore represent one or other sex. Also, as both these systems represent one locus each, and have effective population sizes ¼ that of autosomal loci, results achieved with them are more likely to be affected by stochastic error, and therefore less representative of the entire genome. Furthermore, it is in the autosome that most genes reside and therefore it is likely to be here that susceptibility loci for complex disorders will be found. A comprehensive analysis of the Antioquian autosome is therefore valuable and has been undertaken here, along with four parental populations with diverse demographic histories. Results may not only be informative with regard to the evolutionary history of the population but may also have practical consequences for studies which attempt to use Antioquia to identify genes important in multifactorial traits.

3.4.1.1 Gene Diversity

The mean gene diversities over 17 regions, based on a total of 243 SNPs, showed Antioquia to be the most genetically diverse population; higher than Spain and Beni. This is interesting as it makes Antioquia more diverse than two old, outbred populations: the Beni and the Spanish (a parental population to Antioquia). What's more, when the mean haplotype diversity was calculated (using four regions of high LD), using a sub-set of the original markers based on high minor allele frequencies in all study populations, Antioquia again is more diverse than Spain. The high Antioquian diversity likely reflects admixture and suggests the sustained presence of an Amerindian and African genetic component to the present day Antioquian autosomal genome.

The low diversity in Beni, calculated here from SNP subset 1, was not expected as researchers have previously shown African populations to contain the most diversity of all the world's populations, using autosomal, mitochondrial and Y chromosome DNA (Bowcock et al., 1994; Cann et al., 1987; Hammer, 1995; Jorde et al., 2000; Tishkoff et al., 1996). This has been explained by the antiquity of African populations, consistent with the African origins of modern humans (Cann et al., 1987; Jorde et al., 2000; Tishkoff et al., 1996). Alternatively, and not necessarily a mutually exclusive phenomenon, it may represent an early expansion in Africa before the migration of ancestral non-African populations (Harpending et al., 1998; Stoneking et al., 1997).

A possible reason for the unexpectedly low result for Beni based on all markers is an ascertainment bias for the markers of this dataset. A large proportion of SNPs were selected on the basis of high minor allele frequencies in a population of European descent, and other sources have shown that many markers have confirmed large frequency differences between European and African populations (Collins-Schramm et al., 2002). Therefore the selected markers here could be prone to an ascertainment bias that inflates gene diversities in European derived populations compared to African populations (Mountain and Cavalli-Sforza, 1994; Wakeley et al., 2001). When diversities were calculated from the ML haplotypes, where markers were selected on universal criteria and therefore less sensitive to this type of bias, Beni was the most diverse population, consistent with theoretical expectations.

On the other hand, diversity values obtained for Ticuna were expected. Ticuna is situated in the south-east of Colombia and is a relatively small, isolated population

of constant size. Accordingly, it proved to be the least diverse of all populations for both individual markers and LD haplotypes.

3.4.1.2 Genetic Structure

Antioquia has a complex history of admixture resulting from the arrival of the Spanish colonisers in the 16th century (and eventually their African slaves), the depletion of the local Native American population and possibly Spanish male orientated directional mating which likely occurred during the entire period of Spanish occupation. Studies, which have mainly focused on uniparental inherited genetic systems (Carvajal-Carmona et al., 2003; Carvajal-Carmona et al., 2000) have illustrated these founding events. Their effect on the genetic identity of contemporary Antioquia at the level of the autosome has been less well investigated. One reason for this is that autosomal recombination, absent from mtDNA and the NRY, can greatly confound interpretation of haplotype-based analysis. Here, I have assessed structure between Antioquia and the four parental populations using 425 bi-allelic markers spanning seventeen genomic regions. Results from pairwise population F_{ST} analysis have shown Antioquia to be genetically similar to Spain, and these are the only two populations not significantly differentiated. Additonally, there is also substantial consistency in LD haplotype distribution between Antioquia and Spain. My results, therefore, demonstrate the majority of Antioquian autosomal DNA is of Spanish origin and confirms a large contribution by the Spanish founding population. This is confirmed by admixture analysis of this data which shows estimated ancestry of Antioquia to be 63% European, 32% Native American and 5% African (Ruiz-Linares, unpublished data). These results agree with a previous study on the genetic admixture of Antioquia based on blood-group data which showed around 70% of Antioquian DNA is European (Sandoval et al., 1993). These findings are not in disagreement with the theory of Spanish male orientated directional mating (Carvajal-Carmona et al., 2000).

The relationship of Antioquia to the other populations is illustrated in the population phylogeny which considered the pairwise F_{ST}s as genetic distances (figure 3.3.1). The most parsimonious position of Antioquia is clearly shown to be intermediate to the other populations. For example, Antioquia is closer to both Spain and the African Beni (themselves distant populations) compared to other populations. This is consistent with the ancestral admixture of Antioquia. Results from haplotype

sharing, generated by only including genomic regions with strong LD and low recombination, repeatedly show a trend in support of these findings; haplotypes that are high in frequency in Spain and low in Ticuna/Chipewayan are also high, but not as frequent, in Antioquia. In fact, Antioquia repeatedly shares haplotypes that are common in Spain, Beni or the Native Americans. This work also shows that newly found bi-allelic markers may be as useful for determining genetic relationships of populations as more widely-used, established markers.

3.4.1.3 Linkage Disequilibrium

Several reported events in Antioquian demographic history could theoretically increase LD in this population. These include: it is a relatively young population; it was founded by an admixture event between two genetically distant populations; a population contraction occurred during foundation as a result of the Spanish colonisers representing a small subset of the total Spanish population and their genocidal effect on local native populations; and the small, constant size of the pre-Columbian Amerindian ancestral population.

In contrast to theoretical expectations my results do not show LD in Antioquia to be particularly strong. In particular, a substantial increase in LD in Antioquia compared to the Spanish is expected but was not seen here. These findings could reflect the insensitivity of my test to detect differences in LD, based on marker selection. The SNPs used in the LD analysis were selected on the basis of high minor allele frequency and showed substantially higher mean heterozygosity than the full marker set (i.e. the marker set used for pairwise F_{ST}s and genetic diversities). This means they are unlikely to be able to detect LD generated by diversity reducing mechanisms such as genetic drift, and drift may have generated LD in Antioquia during the founding of the population. The markers used are also not well suited to detecting LD generated by admixture, as no consideration was given to allele frequencies in parental populations. It is alleles with most divergent frequencies in parental populations that give rise to the strongest, admixture generated LD (Collins-Schramm et al., 2002; Zhu et al., 2005). Additionally, common SNPs are likely to be old and therefore a lot of time will have passed for LD to break down, meaning overall values may be reduced.

3.4.1.4 Implications for Antioquia in Gene Mapping

Genetically homogeneous populations remain a popular choice of study sample for complex trait gene mapping (Escamilla, 2001; Varilo and Peltonen, 2004; Wright et al., 1999). Such populations increase the likelihood that all individuals affected with a disease possess the same causal gene. If the population derived from a small number of founders then there is also a higher chance that genomic regions possessing disease genes are identical by descent, thereby increasing the power of LD association studies.

Antioquia may be regarded as a population isolate for several reasons: there was strong differentiation in this region of pre-Columbian Colombia, with tribes separated by the mountainous terrain, leading to low amounts of gene flow and genetic diversity; the present day population was borne from mostly male Spanish colonisers representing a relatively small Spanish subpopulation, and mostly female Native Americans whose effective population size had been drastically reduced by the colonisers; Antioquia is a relatively young population with the major founding event occurring during the late 17^{th} century; since its founding immigration into the province of Antioquia has been limited (Alvarez, 1996).

Despite these factors, evidence presented here does not demonstrate a unique genetic identity to Antioquia, or find it to be particularly genetically homogeneous. My results suggest that the autosomal genome of present day Antioquia is essentially Spanish. What's more, compared to Spain (an old, outbred population), Antioquia are more genetically diverse and show no obvious increase in LD. These findings likely reflect admixture generated gene diversity and perhaps a substantial level of diversity within the Spanish founding population.

Rather than devalue Antioquia as a useful population isolate, however, the genetic characteristics of Antioquia as presented here may be why this population could be very important in gene mapping. As the Spanish contributors represent a small sub-population of Spain, Antioquia is effectively a Spanish population isolate in terms of its autosome and Y chromosome, with a comparatively small number of founders. Complex autosomal disorders, for which many causal mutations exist, may arise in Antioquia due to a subset of these mutations, thereby increasing the chances of finding disease genes due to the decrease in disease heterogeneity. This would apply to disease loci located in the Spanish derived chromosomes at least, and could be applied equally to Native American derived chromosomes if knowledge of disease

epidemiology in Antioquia's Native American ancestral population was available. An increase in probability of chromosome regions that give propensity to disease are indentical by descent may also be achieved, helped by a minimal level of immigration. This leads to a greater efficiency of LD mapping and indirect association analysis (Freimer et al., 1997; Varilo and Peltonen, 2004). Due to only moderate levels of autosomal LD, Antioquia may be more suited to fine-mapping and candidate gene association as opposed to genome-wide association analyses based on long-range LD between a marker and disease allele.

As suggested, Antioquia may be valuable in genetic admixture mapping. Admixture mapping attempts to reveal over-representation of one of the ancestral population's chromosome regions in disease cases, and has recently been used to successfully map regions for hypertension (Zhu et al., 2005). The success of such an approach depends largely on gene effect, and may be more effective when incidence of diseases and frequencies of disease alleles in ancestral populations are very different (Darvasi and Shifman, 2005; Patterson et al., 2004). These parameters need to be determined for Antioquia, although the second most differentiated pair of populations included Ticuna (a Colombian Native American population) and Spain, suggesting the development of this approach in Antioquia may be useful. However, given the majority of autosomal Antioquian DNA is of Spanish origin the efficiency of this approach is not obvious and may only be viable for a small number of disorders.

3.4.2 Genomic Variation in LD

When the strength of LD in each population at various distance bins is assessed in more detail, large standard deviations are seen at all distances (figures 3.3.2 to 3.3.6). It may be that the large variation seen for the greater distance bins is caused partly by the relatively small number of measurements, for e.g., ten for 80kb and nine for 160kb. However, standard deviations are large for both the 20kb and 40kb bins, both of which have considerably more measurements. As the plotted data points are means over seventeen unlinked genes it is probable the observed large standard deviations are representative of substantial genome wide variation in the extent of LD. Analysis of LD within in each region confirms this. For example

some genes have low levels of LD such as *LAMB1*, *KCNA1*, *PCI* and *IL17R*; whereas other genes show extensive LD including *DDR1*, *WASL*, *NF1* and *HCF2*. These patterns are generally well preserved across all populations. Together these findings agree with the gene-specific LD patterns demonstrated in Reich *et al*'s (2001) original study in a Utah population of European descent (Reich et al., 2001).

Stochastic factors such as different gene histories at different gene regions can give rise to variation in levels of genome-wide LD; and may come about by various demographic events. However, the fact that the demographic histories of the five study populations here are markedly different argues against this; for stochastic or demographic factors to explain my results their effects must have acted on the genomes of a population ancestral to all five of my study populations and persisted until present day. An alternative explanation is that some form of selection has acted to keep certain gene regions intact from reshuffling by recombination in humans (and possibly other species). The idea that rates of recombination may vary across the genome and recombination may be localised into hotspots has been proposed and confirmed in recent years (Daly et al., 2001; Jeffreys et al., 2001; Jeffreys et al., 2005). Furthermore, a strong correlation between LD and local recombination rate has been shown (Reich et al., 2001). This would separate the human genome into regions of high and low LD with *DDR1*, *WASL*, *NF1* and *HCF2* residing in LD islands, and explains well the patterns of LD described here. However, such a model is in contrast to the conclusions of Reich *et al* (2001) who did not use the presence of recombination hotspots to explain the genome-wide distribution of LD, but rather a population bottleneck 27,000 to 53,000 years ago was shown to be the most likely cause of long-range LD in northern Europeans. It should be noted that similarity in LD patterns between populations in my study may be overestimated as a result of the marker selection which had a bias for more frequent minor allele SNPs; a necessary measure to facilitate cross population comparisons. Nonetheless, short-range LD is shared between five populations with distinct demographic histories.

3.4.3 Chipewayan

While assessing the genetic relationships of Antioquia with parental populations, the genetics of the Chipewayan have been exposed in more detail. The Chipewayan are part of the NaDene language group whose ancestors are believed to

have been part of an early migration from Asia into North America, across the Beringian land bridge (Merriwether and Ferrell, 1996; Merriwether et al., 1996). Since then the Chipewayan would have become a separate population remaining relatively isolated with little gene flow with other populations. Genetic structure analysis from pairwise population F_{ST}s shows that the Spanish and Antioquia are closer to the Chipewayan than the Ticuna. Furthermore, haplotypes that are common in Spain and Antioquia are more common in the Chipewayan than the Ticuna, such as NF4 and HCF5. It would appear, therefore, that the Chipewayan share a modest level of genetic ancestry with Europe.

These results could be explained by the 'Spanish' haplotypes migrating to both Spain and North America from Africa or Asia, but not reaching South America; in agreement with evidence, based on the Y chromosome, of at least two separate migrations by modern humans into the Americas (Bortolini et al., 2003; Lell et al., 2002; Ruiz-Linares et al., 1999). Alternatively, my results are also consistent with a more controversial idea suggesting that, during the European colonisation of North America, European and native North American population mixing occurred (Tarazona-Santos and Santos, 2002). Determining the frequency of the 'Spanish' haplotypes in populations from North-East Asia (in particular middle Siberia and the Lower Amur regions) would help to resolve this. For example, if Siberian populations were found to have these 'Spanish' haplotypes at frequencies intermediate between Spain and the Chipewayan, then this would argue against European admixture. If these haplotypes are absent then admixture is supported, and would indicate that genetic mixing of Europeans and Native Americans was relatively frequent in North, Central and South America (Bortolini et al., 2003; Carvajal-Carmona et al., 2000; Green et al., 2000; Merriwether et al., 1997; Tarazona-Santos and Santos, 2002). As well as important for historical records, this could have implications for disease-gene mapping as more admixed populations may exist in the Americas than previously thought: population admixture can be effectively exploited to find disease genes (Darvasi and Shifman, 2005; Zhu et al., 2005).

3.4.4 Global Colonisation

3.4.4.1 Origins of Modern Humans

The evolutionary history of modern humans remains a keenly debated topic. While an African origin of AMH around 100KYA is rarely disputed, the number of colonising migrations is. The most widely held view is that one major Upper Palaeolithic expansion gave rise to all contemporary genetic diversity (Bowcock et al., 1994; Cann et al., 1987; Underhill et al., 2000). However, recent work by Zietkiewicz *et al* (2003) provides contrary findings. They discovered a small, old haplogroup of the dystrophin gene at high frequencies in the Americas and Europe, but rare in African populations. They concluded that this haplogroup represented a lineage that was not part of the major out-of-Africa expansion, but instead formed part of a secondary contribution to contemporary world-wide human genetic diversity. It could either have originated in Africa before the Upper Palaeolithic expansion and was carried out of Africa in a minor migration where it flourished, but subsequently waned in Africa; or this lineage may represent a previously undocumented non-African contribution to the human gene pool (Zietkiewicz et al., 2003). Although these conclusions were based on data from one gene, and may therefore be subject to the effects of stochastic events such as drift (which may lead to erroneous interpretation), their observation is an interesting one.

Here, I have identified an *NF1* haplogroup, whose sole member is haplotype NF9, which is common in Ticuna, present at low frequencies in the other South American populations, but completely absent from Beni and Spain (figure 3.3.22). Although traces of this haplotype would be expected in either Beni or Spain if one major out-of-Africa migration occurred, haplotype NF9 may have arisen due to recombination or mutation of a haplotype that left Africa but did not reach Europe. However, according to haplotype network analysis five mutation steps separate NF9 and the two haplotypes most common in Beni: NF1 and NF2 (data not shown). Eight mutation steps separate NF4 (accounting for 80.6% of Spanish haplotypes but only 6.5% for Beni) from both NF1 and NF2. NF9 and the haplogroup containing NF4 and NF11 (virtually absent from Beni) are closely related to one another and share a non-African allele at marker 2 and an allele at very low African frequency at marker 5 of these haplotypes. This may suggest a separate origin for these haplotypes compared to

Beni and could represent either a non-African contribution to the current human gene pool, or an early African contribution that has since disappeared in the Beni.

My data are not inconsistent with the findings of Zietkiewicz *et al* (2003); however it is important to determine the origin of haplotypes NF9, NF4 and NF11. Increasing the numbers and world-wide distributions of populations sampled will be very useful as my study has been limited to five populations, none of which originate in Asia. In particular only one African population is included, from Nigeria in West Africa. It has been long known that Africa harbours a large proportion of human genetic diversity (Jorde et al., 2000; Stoneking et al., 1997), meaning the Beni are not likely to represent well the genetic make-up of the Palaeolithic emigrants. As such, African populations from eastern, southern and northern parts are required for a more comprehensive African representation, especially since East Africa is likely to be the site from which many non-African populations migrated. Estimating the antiquity of haplotypes NF9, NF4 and NF11 would also be very informative, made feasible by constructing an ancestral haplotype derived from non-human primate allelic states. If the non-African haplotypes evolved prior to the Upper Palaeolithic expansion further support would be given to a second origin of contemporary human genetic diversity. Dependent on such evidence, the 'Out-of-Africa' theory of modern human evolution may need revised.

3.4.4.2 Population Expansions

Contrasting opinions exist regarding the demographic and biogeographic patterns of modern humans leading to our successful colonisation of the globe. Research has generated variable results including: a single expansion in Africa prior to migration to the rest of the world (Excoffier and Schneider, 1999; Jorde et al., 1997; Zhivotovsky et al., 2003); multiple expansions of modern humans from Africa (Templeton, 2002); Pleistocene expansions in all major continental groups (Excoffier and Schneider, 1999; Templeton, 2002; Zhivotovsky et al., 2003), expansion in Africa prior to ancestral migration, but none elsewhere (Reich and Goldstein, 1998); expansions outside Africa only (Kimmel et al., 1998; Rogers and Harpending, 1992); and a bottleneck prior to expansion in Europeans (Reich et al., 2001). Reasons for such disagreement of ancestral population growth may be that global colonisation has involved a complex series of evolutionary factors, and tests have not been powerful or sensitive enough to accurately identify each separate factor. Analysis of results is

further compounded if selection has acted upon the test loci, or if variable mutation rates have been used in calculations of divergence times which is likely as robust estimates are hard to estimate in humans(Schneider and Excoffier, 1999).

Over the past fifteen years several techniques have been developed to detect genetic traces of historic demographic events in ancestral populations. One approach is to look at the number of polymorphic sites between two sequences or haplotypes within a population. These are also referred to as pairwise differences and their frequency distribution is known as a mismatch distribution; dependent on mutation rate, population size and generation number (Rogers and Harpending, 1992). Rogers and Harpending (1992) proposed that, under an infinite-sites model, the shapes of these distributions within populations illustrate patterns of demographic growth. For example, a unimodal distribution signifies an expansion of exponential growth, and a similar pattern may also arise as a result of a population bottleneck. Multimodality, on the other hand, represents population stationarity (Rogers and Harpending, 1992). This theory has since been reinforced, although it is now believed that unimodal distributions may also be caused by selective sweeps and mutation rate heterogeneity (ArisBrosou and Excoffier, 1996; Harpending et al., 1998; Schneider and Excoffier, 1999).

Tajima's D is another measure of a population's deviation from a neutral situation. This is based on a comparison of genetic diversity measured under the infinite-alleles model, π or θ_π (the mean number of pairwise sites), to genetic diversity measured under the infinite-sites model, θ_S (based on the number of segregating sites). Tajima exploited the fact that selection may affect these values in separate ways. For example, if selection acts upon a deleterious allele it will reduce the frequency of that allele, but the allele will most likely remain in the population at low frequencies for a longer period of time. θ_S is independent of allele frequencies and is not affected by a reduction in the numbers of alleles so long as they persist in a population; whereas θ_π is directly dependent on allele frequencies. When θ_S is taken away from θ_π a negative value is therefore taken as evidence for directional selection (Tajima, 1989). A similar influence on D is expected after a large, recent population expansion as an expansion will increase the number of rare alleles in a population but not enough time will have past for them to reach large frequencies (Hartl and Clark, 1997). Significantly positive D values, on the other hand, may indicate balancing selection or population differentiation (Bamshad et al., 2002).

Fu's F_S test of selective neutrality, also based on the infinite-sites model, compares the number of alleles in the observed dataset to that expected in a neutral population, using the observed numbers of pairwise differences to make the neutral expectations. This test is particularly sensitive to population expansions, indicated by large negative values; an anomaly of this test is that a p value of 0.02 is required to achieve significance at the 5% level (Fu, 1997).

From my analyses, a unimodal mismatch distribution was observed for Chipewayan in three out of the four high LD regions, providing strong evidence for recent substantial growth for this population. This result supports other studies on the peopling of the New World based on mtDNA (Bonatto and Salzano, 1997). However, some research using autosomal (Zhivotovsky et al., 2003) and mtDNA (Excoffier and Schneider, 1999) has not been able to show expansions in populations from America, including hunter-gatherers.

One gene, *DDR1*, generated non-neutral distributions in all five populations (figures 3.3.34 to 3.3.38), and may provide insights into the expansion of modern man out of Africa. The mismatch curves for this gene all peak at 8 differences in Chipewayan, Ticuna and Antioquia, whereas the Beni curve peaks at 3 differences. Spain is interesting in that its distribution is bi-modal: peaks exist at 3 and 8 differences (figure 3.3.36). If the peak at 3 pairwise differences in Beni has arisen due to a demographic event it would appear a population bottleneck or expansion occurred in West Africa before, and perhaps leading to, expansion into southern Europe. A population expansion or reduction outside West Africa is likely responsible for the peak at 8 differences: a mismatch distribution present in Europe and the New World but absent from West Africa. This theory is somewhat supported by the population frequency distributions of the *DDR1* haplotypes (figures 3.3.11 to 3.3.15), where DR5 and DR8 are common in all populations except for Beni, in which DR1 is prevalent. Mismatch distribution results for this gene agree with theories that a major change in population size occurred in Africa prior to expansion into Europe and before world-wide colonisation out of Africa. The mismatch distribution shared between Beni and Spain, and not the other populations studied here, suggests that separate events describe the variation seen in Beni compared to the New World. However the evidence is not strong and, in particular, Tajima's D values for *DDR1* do not support a deviation from neutrality, all values are positive.

Additionally, an interesting distribution of *NF1* haplotypes in Spain was revealed and may require further analysis. A strong predominance of only a few *NF1* haplotypes in the Spanish was found; NF4 accounted for 80.6% of all haplotypes and NF2 for 16.4% (figure 3.3.22). There may be several reasons for this result in such a large and relatively old population. It could be that in the other populations diversity has been inflated due to either age (for Beni) or admixture and that the level seen in Spain represents what would be expected for this genomic region based on mutation and recombination rates alone. This does not seem likely; modern man reached Spain 20-30 thousand years before America and, of the New World populations studied here, only Antioquia has a documented admixture event. Although genetic drift could explain decreased diversity levels in this Spanish population, drift is more likely to occur in small, isolated populations. Additionally, there is no evidence for low variation elsewhere, either in terms of genetic diversities or by other haplotype diversities. The best explanation for the homogenous *NF1* distribution may simply be a unique history for this gene in Spain, or rather in the region of Spain from where these samples derived. Alternatively, the *NF1* gene encodes the neurofibromin protein, which is involved in the ras signal transduction pathway and is involved in the neurofibromatosis disorder, so there is a possibility that a selective constraint exits for this gene in Spain; however, the rarity of this disease (with prevalences as low as 1 in 5000) suggests that effects on disease loci will not be observed at the population level. Also the mismatch distribution is not statistically unimodal for this gene and Tajima's D was low, but not statistically significant (see table 3.3.11). Further description of *NF1* haplotype distribution and more stringent tests of selection are required to investigate the potentially interesting trend observed here.

It is important to note for all these analyses that Beni may not characterise the founding African population that gave rise to world-wide genetic variation. Interpretation of these results would be substantially aided by increasing the number of populations studied, and in particular more African populations from different regions including East Africa could be very beneficial. Furthermore, more than one factor may generate patterns of expansion according to the neutrality test used, confounding interpretations. For example, unimodal mismatch distributions combined with positive D values have been explained elsewhere as reflecting heterogeneous mutation rates, rather than population expansions (ArisBrosou and Excoffier, 1996). If only a few sites are able to mutate (a consequence of unequal mutation rates),

mismatch distributions can also overestimate the effects of a population expansion. As several factors can influence the distribution of pairwise nucleotide differences including mutation rates (and their uniformity), magnitude of expansion, time since expansion and initial population sizes, further analysis is needed before conclusions can be drawn from these results.

3.4.5 F_{ST} Distribution

Using genetic distance as a means of detecting selection was first proposed several decades ago (Cavalli-Sforza, 1966), but it has only recently been made feasible due to advances in available human genome sequence, and identification of useful polymorphic markers within it. The underlying principle is that directional selection should increase genetic distances between populations exposed to different environmental conditions, thereby generating large F_{ST}s. Conversely, extremely small F_{ST}s will be generated at loci under balancing selection, as a result of populations converging to a common genotype. The significance of these extreme F_{ST}s can be obtained by comparing them to a simulated distribution assuming neutrality (Bowcock et al., 1991; Lewontin and Krakauer, 1973), or by using the empirical distribution (Akey et al., 2002; Kayser et al., 2003).

Here, global F_{ST} estimates for 395 SNPs were generated and a mean value of 0.1041 was obtained which lies within a range of previous estimates of global F_{ST}s (Barbujani et al., 1997; Bowcock et al., 1991; Jorde et al., 2000; Romualdi et al., 2002). Four markers gave F_{ST} values ≥ 0.4, which represent 1.01% of the distribution, and seventeen markers gave F_{ST} values of 0, which represent 4.3% of the total distribution; these tail proportions are less than in a previous study by Akey *et al* (2002) that used 26,530 genome-wide gene associated SNPs.

In Akey *et al*'s (2002) study the empirical F_{ST} distribution was used to detect markers under the influence of directional selection and an F_{ST} was regarded to signify directional selection if above 0.45, as this generated a p value of 0.026 based on the empirical distribution; whereas genes were considered to be under balancing selection if they contained two F_{ST}s equal to 0 and one ≤ 0.005, which were estimated to represent a significance level of 0.03 based on simulations. Using these criteria only one gene shows evidence of directional selection: *DDR1* (discoidin domain receptor family, member 1), which encodes a tyrosine kinase receptor involved in

development and differentiation of neurons. On closer inspection it seems that it is differences in allele frequencies between the Native American populations and the rest of the world that is responsible for the high overall F_{ST} for this marker (table 3.3.12). If the criteria is broadened slightly *SCYA2* (small inducible cytokine a2) may also show evidence of evolutionary forces, as this gene contains three SNPs with F_{ST}s higher than 0.4 and all lie within 2.5kb of one another. This is good evidence for selection as Akey *et al* (2002) showed that adaptive hitch-hiking or background selection is the most likely cause for a correlation between high F_{ST} and physical distance. *SCYA2* is a member of the small inducible gene (SIG) family and plays a role in the conscription of monocytes to wounds and infection sites. Directional selection could act on this gene if different populations are exposed to different infection related diseases throughout their history. Only one gene showed evidence of balancing selection: *KCNA1* (potassium channel, shaker subfamily, member 1), encodes a voltage-gated ion channel involved in modulating membrane potential and the polarisation of neurons. A species-wide selective constraint on such a gene in species with an advanced central nervous system, such as *Homo sapiens*, is easily envisaged.

Overall, proportions of F_{ST} markers that lie in the extremes of the overall distribution are less here than in previous studies (Akey et al., 2002; Kayser et al., 2003). This may suggest a lack of detectable evolutionary forces acting on these seventeen genes. This trend is also likely strongly influenced by the substantially fewer markers being used here, and representation of only seventeen gene regions. In contrast, three out of seventeen genes have potentially shown evidence of selection and this percentage is high compared to other studies. Comparisons of results to previous studies are made difficult by the relatively small sample sizes in this study. Nonetheless, results here may be important in the understanding of how selection has shaped contemporary genetic diversity and, importantly, may have highlighted superior functional candidate genes for common complex disorders.

CHAPTER 4: ASSOCIATION ANALYSIS OF *SLC6A4* WITH BIPOLAR AFFECTIVE DISORDER IN THE ANTIOQUIA POPULATION ISOLATE

CHAPTER 4: ASSOCIATION ANALYSIS OF *SLC6A4* WITH BIPOLAR AFFECTIVE DISORDER IN THE ANTIOQUIA POPULATION ISOLATE

4.1 Introduction

4.1.1 Psychiatric Disorders

Psychiatric disorders are a class of complex trait concerned with abnormal behaviour and mood. Underlying mechanisms remain elusive due to their polygenic and heterogeneous nature, consistent with other complex traits. Additionally, due to a lack of biological markers, their identification and monitoring through developmental stages relies on individual assessment of patients by doctors that may not use the same official criteria sets (Bearden et al., 2004; Reus and Freimer, 1997).

Although not fully understood, the environment is expected to play a significant role (Duffy et al., 2000; Owen et al., 2000). For example, negative life experiences (such as death of loved ones) are believed to significantly influence depression related disorders; and there is evidence for higher rates of schizophrenia and conduct disorders in unstable family environments (Cadoret et al., 1995; Tienari et al., 1994).

A strong genetic influence has also been indicated for some neuropsychiatric disorders from studies on family prevalence and estimations of heritability (Cardno et al., 1999; Jablensky, 2000; Kendler et al., 1995a; Kendler et al., 1995b; Sawa and Snyder, 2002). However, poor understanding of the gene-environment interaction compounds molecular characterisation, as does evidence that suggests perception of the environment is genetically controlled; so both disease susceptibility and environment sensitivity genes may be required (Kendler et al., 1993).

These factors make studying psychiatric disorders particularly challenging. Initial, putative genetic associations are not well replicated in later studies. Consequently, specific susceptibility loci remain unidentified for many disorders. To improve association efforts consideration could be made of the characteristics of disease relevant to its genetic epidemiology such as; studying variants that have a higher likelihood of being causal, using markers in indirect approaches that are in strong LD with functional parts of genes, and a more careful selection of candidate

genes based on a strong possibility of involvement in the biology or aetiology of the disorder (Tabor et al., 2002).

4.1.2 BPAD

Affective (or mood) disorders are a group of psychiatric disorders that involve abnormal mood regulation including major depression, bipolar affective (or mood) disorder and dysthymic disorder (Medicine, 2003; Reus and Freimer, 1997). Although each of these mood disorders is distinct, there can be considerable overlap between phenotypes (Escamilla et al., 1997). BPAD (bipolar affective disorder) is characterised by recurrent episodes of mania or hypomania, major depression and mixed episodes. According to the American Psychiatric Association a major depression episode consists of at least two weeks in which there is a loss of pleasure in almost all activities. A manic episode is described as a period of time when a patient's mood is abnormally high, confused or irritable whilst hypomania, as the name suggests, is a less severe form of mania (Association, 1994). In a mixed episode both manic and depressed symptoms co-occur within a one week period (Association, 1994). BPAD generally occurs equally in men and women and has a peak age of onset between fifteen and twenty-four years, with a population prevalence between 0.5 and 1.6% (Kessler et al., 1994; MacKinnon et al., 1997; Muller-Oerlinghausen et al., 2002; Robins et al., 1984). The annual cost of this disorder to the UK government was estimated at £2 billion sterling for the years 1999 and 2000 (Das Gupta and Guest, 2002).

Several classes of phenotype have been identified with BPI regarded as the most severe, characterised by mania or mixed episodes recurring with major depression. Additionally, rapid cycling can occur with BPI (mostly in women) in which more than four episodes occur within a twelve month period. Also associated are a 20-30% suicide risk and a high rate of co-morbidity with other disorders, such as hyperactivity attention deficiency and substance abuse (Fagiolini et al., 2004; MacKinnon et al., 1997; Muller-Oerlinghausen et al., 2002). BPII consists of mixed episodes of major depression and hypomania whereas cyclothymia, a third form, involves minor depression and hypomanic episodes.

Due to its severity, of BPI in particular, a lot of research has focused on BPAD but a definitive cause has not yet been identified. Identifying the determinants of

BPAD is a complicated issue due to the complex nature of the disease, phenotype overlap with other mood disorders and the potentially subjective assessment by doctors and psychiatrists. Additionally, the different phenotype classes may not be caused by the same susceptibility factors (Spence et al., 1995). In spite of this, observations of family prevalence show a strong genetic contribution. Investigations have shown a concordance rate of 50-70% for monozygotic twins and 13% for dizygotic twins in BPI, substantially larger than the general population prevalence (Gershon et al., 1990; MacKinnon et al., 1997; McGuffin and Katz, 1989). Furthermore, the relatively high heritability estimate for BPI of approximately 80% also supports a strong genetic contribution (Kendler et al., 1995a). In accordance, considerable work has been done to more fully understand the genetic basis of the disease.

4.1.3 Mapping a Disease Gene

Although a strong genetic component is proven, identifying a single causal gene or group of genes has not been achieved for bipolar affective disorder, despite the efforts of many researchers. Reasons are likely to include gene effects being too modest, the rarity of susceptibility alleles and a high level of disease heterogeneity. These factors increase the difficulty of finding genes and repeating positive findings in independent samples (Bearden et al., 2004; Zondervan and Cardon, 2004). As a result it is believed that association studies may be well suited to finding genes for psychiatric disorders such as bipolar since, compared to linkage analysis, these studies are sensitive to genes of modest effect and consideration of environmental factors in the study design is more feasible (Bell and Taylor, 1997; Hirschhorn and Daly, 2005). Family-based association studies have the added benefit of removing the risk of obtaining false-negative results as a result of population stratification.

4.1.3.1 Candidate Gene Selection

A critical first step to an association study is to identify a suitable candidate gene based on the relevance of its function to the disorder in question. There are two main approaches to this. One is to use positive results from linkage analysis to highlight a chromosomal region for further screening. If a gene is found within a

highlighted region that is likely to influence the phenotype, then it would warrant mutation screening to find susceptibility alleles.

Results indicate many different genomic regions may be important in BPAD (see table 4.1.1). Before these results can be deemed reliable however they require independent replication, and only a few achieve this, including 12q24 (Degn et al., 2001; Ewald et al., 1998; Morissette et al., 1999), 18p11 (Escamilla et al., 1999), 18q22 (Fallin et al., 2004; Freimer et al., 1996; Garner et al., 2001; McInnes et al., 1996; McMahon et al., 2001), 21q22 (Aita et al., 1999; Straub et al., 1994) and Xq28 (Baron et al., 1994; Bocchetta et al., 1994). Interestingly, a genome scan meta-analysis performed by Segurado and colleagues, using only studies with at least 20 affected cases, supported nominal linkage to 8q24, 9p21, 10q11-22, 14q24-32 and 18p11-q22 (Seguardo et al., 2003). Other studies have shown lack of linkage with these regions such as that done by Knowles *et al* (1998) who used pedigrees from different populations and found no linkage with chromosome 18 (Knowles et al., 1998). Existence of linkage to the X chromosome in particular has been keenly debated since it was first suggested in 1969 (Bocchetta et al., 1994; Reich et al., 1969).

Table 4.1.1. Summary of positive linkage analyses for BPAD.

CHROMOSOME REGION	AUTHOR
4p	(Blackwood et al., 1996)
5p	(Kelsoe et al., 1996)
5q	(Garner et al., 2001)
6 (non-significant)	(Ginns et al., 1996)
11p	(McInnes et al., 1996)
12q	(Degn et al., 2001; Ewald et al., 1998; Morissette et al., 1999)
13q	(Badenhop et al., 2001)
15 (non-significant)	(Ginns et al., 1996)
16p	(Ewald et al., 1995)
18 pericentromeric	(Berrettini et al., 1994)
18p	(Escamilla et al., 1999; McInnes et al., 2001; Stine et al., 1995)
18q	(Fallin et al., 2004; Freimer et al., 1996; Garner et al., 2001; McInnes et al., 1996; McMahon et al., 2001; Stine et al., 1995)
21q	(Aita et al., 1999; Straub et al., 1994)
Xq	(Baron et al., 1987)

A potential problem with this approach is that complex traits may be caused by convoluted interactions between many genes, and these genetic pathways are generally poorly understood. Therefore, a good candidate gene located within a linked region may easily be overlooked. Also, linkage studies are not well suited for finding genes of modest effect, increasing the chances of obtaining false-negative results (Hirschhorn and Daly, 2005).

A second approach of candidate gene selection is to predict more fully genes or groups of genes that are likely to be important in expressing a phenotype. Genes can be selected on the basis of their biological function, by extrapolating information collected from studies in other related psychiatric traits and analysing the targets for drugs that are used to treat the disorder. Screening for mutation or variants can then be done in an attempt to find an association. Tested variation may lie in functionally important gene regions and be directly causal, or may be distant from functional sequence but in linkage disequilibrium with the disease causing allele. A summary of genes associated using this approach is presented in table 4.1.2.

The genes of the serotonin neurotransmission pathway pose good functional candidate genes for bipolar affective disorder for several such reasons: (i) these genes regulate the levels and mediate the biological effect of the serotonin neurotransmitter (5HT) in CNS neuronal synapses, important in growth, development and mood regulation (Frazer and Hensler, 1999); (ii) the serotoninergic gene products are the target for major depression and BPAD drugs, namely the selective serotonin re-uptake inhibitors (SSRIs) which impair serotonin transporter activity, as well as emotion changing recreational drugs such as amphetamines (Little et al., 2003; Serretti et al., 2004; Serretti et al., 2001; Uhl et al., 2000); (iii) clinical analyses have revealed abnormal levels and states of the serotoninergic gene products in patients suffering from bipolar disorder (Drevets et al., 1999; Vawter et al., 2000); (iv) they have been shown to be involved in many behaviours and psychiatric traits both in humans and animal models; for example, a strong association between serotonin transmission and anxiety has been demonstrated in mice (Griebel, 1995; Gross et al., 2002). Therefore, there have been many attempts to associate these genes with BPAD (Bellivier et al., 1998; Collier et al., 1996; Lerer et al., 2001; Lim et al., 1995; Rotondo et al., 2002).

Table 4.1.2. Summary of positive association of functional candidate genes with BPAD.

GENE	CHROMOSOME	AUTHORS
SLC6A4	17q11.1-12	(Battersby et al., 1996; Rotondo et al., 2002)
TPH1	11p14-p15.3	(Bellivier et al., 1998)
MAOA	Xp11.4-11.3	(Lim et al., 1995)
HTR2C	Xq24	(Lerer et al., 2001)
COMT	22q11.2	(Kirov et al., 1999; Li et al., 1997; Rotondo et al., 2002)
TH	11p15.3-p14	(Meloni et al., 1995)
DRD2	11q23	(Massat et al., 2002)
BDNF	11p14.1	(Sklar et al., 2002)

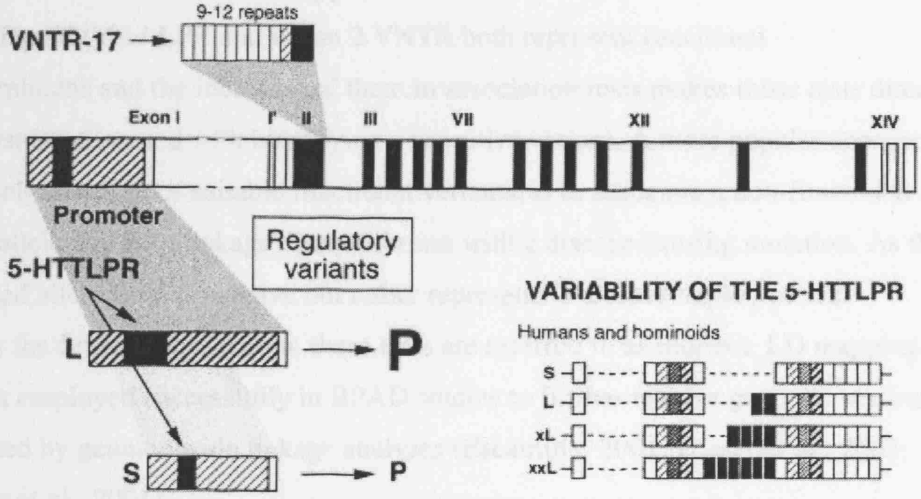
SLC6A4: serotonin transporter; TPH1: tryptophan hydroxylase 1; MAOA: monoamine oxidase A; HTR2C: serotonin receptor 2C; COMT: catechol O-methyltransferase; TH: tyrosine hydroxylase; DRD2: dopamine receptor 2; BDNF: brain derived neurotrophic factor.

One of the serotoninergic genes, *SLC6A4* (human serotonin transporter), has been of particular interest. It maps to chromosome 17 q11.1-q12 (Ramamoorthy et al., 1993), has fourteen exons and spans approximately 40kb. The serotonin transporter mediates the neuronal re-uptake of 5HT via Na+ dependent ion exchange from the synapse, where 5HT stimulates synaptic receptors (Frazer and Hensler, 1999). Therefore, this gene has a direct effect on the levels of synaptic serotonin and consequently synaptic serotonin function.

Interestingly two common *SLC6A4* functional polymorphisms have been documented (see figure 4.1.1). One, situated in intron 2, is a VNTR consisting of nine, ten, eleven or twelve copies of a 17bp repeat unit (Lesch et al., 1994; Ogilvie et al., 1996); and the ten and twelve copy alleles have been shown to have a positive regulatory effect on LacZ expression *in vivo* (MacKenzie and Quinn, 1999). The other, also a VNTR, is located approximately 1kb upstream from the transcription initiation site and was termed the serotonin transporter linked polymorphic region (5HTTLPR, and referred to here as LPR) by Heils *et al* (1996). Initial studies have shown it to consist of 16 repeat units 20-23 bp in length, with the polymorphism generated by a 44bp deletion event involving repeat units 6-8 (Heils et al., 1996). *In vitro* experiments demonstrate a significant effect on expression levels, with long (L) alleles driving expression up to three times more efficiently as short (S) alleles (Heils et al., 1996; Lesch et al., 1996). Additionally, lymphoblast cells homozygous for the L

allele had 1.43 and 1.67 times the concentration of *SLC6A4* mRNA than cells with either 1 or 2 copies of the S allele, respectively and tests on serotonin membrane binding and cell uptake mirrored these trends. Together these results show that the L allele drives expression significantly more than the S allele, and that the relationship between the alleles resembles recessive-dominant.

Figure 4.1.1. The serotonin transorter gene, *SLC6A4*, and functional promoter polymorphisms.



Size of P indicates effect on expression of S and L alleles, with large P more and small P less. XL shows 18 repeat allele and xxL shows 20 repeat unit, both rare in humans. From (Lesch and Mossner, 1998).

Several early studies have supported a role of the serotonin transporter in bipolar affective disorder. Collier and colleagues (1996) used a case-control assay and demonstrated a strong association of VNTR allele 12 with BPAD; later replicated by Kirov *et al* (1999) who used ninety-eight BPI British caucasian trios in a family based association analysis, although a weaker effect was observed (Collier et al., 1996; Kirov et al., 1999). Battersby *et al* (1996) showed that 128 bipolar cases had statistically more of the 9 copy VNTR allele than a group of non-affected controls, with a stronger effect for major depressives (Battersby et al., 1996).

The LPR has also been implicated. In an early study Collier *et al* (1996) demonstrated over-representation of the short allele in a mixed European sample of

bipolar and unipolar patients compared to a control group, although strength of association was relatively weak (p=0.03). An implication of the short allele has since been repeated in other studies including a meta-analysis involving 392 bipolar cases and 739 controls (Furlong et al., 1998; Mynett-Johnson et al., 2000; Rotondo et al., 2002). A more recent meta-analysis further supports a role of the short LPR allele (Lasky-Su et al., 2005). However, many efforts since the initial studies have failed to reproduce positive associations with *SLC6A4*.

4.1.3.2 Indirect Versus Direct Approaches

The *SLC6A4* LPR and intron 2 VNTR both represent functional polymorphisms and the inclusion of them in association tests makes those tests direct, as the test is concerned with identifying a causative variant. A more popular approach, due largely to a lack of suitable functional variants, is to associate a non-functional marker allele that is in linkage disequilibrium with a disease causing mutation. As the associated allele is not causative but rather represents a disease haplotype that includes the functional mutation, these tests are referred to as indirect. LD mapping has been employed successfully in BPAD studies to further resolve genomic regions implicated by genome wide linkage analyses (Escamilla, 2001; Glaser et al., 2005; McInnes et al., 2001).

A more recently developed idea in the field of complex disorder association analysis also exploits LD around candidate genes. This method uses numerous markers to define haplotypes that span gene regions and then attempts to associate multi-marker haplotypes with a disorder. It may be advantageous to use multi-marker haplotypes as more genetic information is included than if single markers are used, although a statistical drawback is the increased number of degrees of freedom (Seltman et al., 2001; Templeton, 1995). This technique has been recommended for psychiatric disorders (Collier and Sham, 1998) and interesting results have been obtained for attention deficit hyperactivity disorder (ADHD) (Curran et al., 2005) and schizophrenia (Li et al., 2004; Shifman et al., 2002). In fact, in the study by Curran *et al* (2005) 13 markers spanning 51.3 kb were used to define 5' and 3' LD blocks around *SLC6A4*. They demonstrated strong association between ADHD and four marker haplotypes in the 5' block. To date this method has not been used extensively for BPAD although one study that used a limited number of markers is encouraging. Mynett-Johnson *et al* (2000) genotyped 3 markers around *SLC6A4*: the LPR, VNTR

and a 3' UTR SNP. They were able to reveal a strong global significant association for two marker haplotypes involving the LPR and the 3' UTR SNP. A more comprehensive analysis of the haplotype structure in *SLC6A4* would be useful to confirm this result and perhaps highlight more specific gene regions.

For this study I have investigated a potential association of *SLC6A4* with bipolar affective disorder using fourteen markers that span approximately 303.4 kb around the gene. Ten SNPs, three microsatellites and the LPR have been included and tested for association both individually, and as successive four marker haplotypes in a 'sliding window' format, to increase the power of analysis. I have focused on individuals diagnosed with a narrow BPI phenotype, reducing the extent of disease heterogeneity. Further, two population isolates have been used: Antioquia and another Latin American admixed population from the Central Valley of Costa Rica (CVCR). The CVCR isolate has previously been shown to be genetically similar to Antioquia (Carvajal-Carmona et al., 2003). This will allow an opportunity for repetition of results within the same study.

An attempt has also been made to define the LD structure around *SLC6A4* in Antioquia and CVCR, to determine whether LD has been separated into blocks in these populations. Block structure can be compared between the populations as a means of assessing genetic similarity, often representative of a shared demographic history. Furthermore, LD block haplotypes can be tested for association to determine whether certain regions of *SLC6A4* give greater predisposition to BPI than others. This will reduce the excessively large number of comparisons (and hence degrees of freedom) a 'sliding window' method produces, whilst retaining a higher amount of genetic information than single marker analyses.

This study represents a highly detailed and comprehensive analysis of *SLC6A4* and its potential role in BPAD. By including several different approaches and two different populations within the same study, every opportunity has been given to identify important alleles and gene regions. The results should add to the contentious issue of *SLC6A4* as a BPAD locus, and help resolve the issue for the Antioquia and CVCR populations. Such information may help the understanding of the molecular mechanisms underlying BPAD, and ultimately improve treatments.

4.2 Methods

4.2.1 Population samples

For this study I analysed DNA from 73 unrelated Antioquian individuals affected with BPI, as diagnosed using DSM-IV criteria (Association, 1994). These patients were identified at the Hospital Universitario San Vicente de Paul, Hospital Mental de Antioquia and at Clinica Sameín, all in the city of Medellín. DNA was available for both parents in 23 instances, and for one parent only in 50 instances. The case parents and grand-parents are also Antioquian.

BPI affecteds from the Central Valley of Costa Rica were also analysed, diagnosed using the Diagnostic Interview for Genetic Studies (Nurnberger et al., 1994). Interviews were independently reviewed by 2 psychiatrists at University of California San Francisco in an attempt to further limit the phenotype range. DNA was available for both parents in 27 instances and for one parent in 53 cases. Parents and grand-parents were all from CVCR.

Prior to this study, individuals were informed and consent was obtained before blood samples were taken and DNA extracted. In general, methods of ascertainment were similar for both populations and are explained more fully in Escamilla *et al* (1996). An attempt was made to keep the size and structure of the two sample populations similar. This was to maximise power and facilitate comparisons between the populations.

4.2.2 Marker Selection

4.2.2.1 SNPs and LPR

Initially, a total of 10 SNPs and the length polymorphic region (LPR) spanning *SLC6A4* on chromosome 17q11.2 were chosen for typing. SNPs were selected using a combination of programs including: Ensembl; the Human Genome Browser (University of California, Santa Cruz); the SNP Consortium; and the Assays-on-Demand software at ABI (see 4.2.4.5). Five intragenic SNPs, three SNPs 3' and three SNPs 5' to the gene were selected. Intragenic SNPs had heterozygosities of at least 0.10.

Flanking SNPs were located approximately 40 kb apart where possible, and starting at 40 kb from the most 5' and 3' intragenic SNPs. In cases where more than one SNP existed at the correct distance, SNPs were selected on the basis of their validation as summarised by dbSNP at NCBI. Although heterozygosity was the preferred selection criterion, most flanking SNPs' allele frequencies had not been characterised at the time of study. Indeed, four flanking SNPs used had not yet been validated.

The *SLC6A4* LPR is a highly heterozygous, well characterised bi-allelic locus situated approximately 1.2 kb upstream from the transcription start site. This marker was of particular interest as its long and short alleles have already been shown to have significant effects on gene expression (Heils et al., 1996).

4.2.2.2 STRs

As genotyping data was collected it became clear that the 5' flanking SNPs were not very informative due to low levels of heterozygosity. To increase information levels on this side of the gene, as well as further define haplotypes around the LPR showing potentially significant associations, three short tandem repeat loci (STRs, microsatellites) were added to the analysis.

The three microsatellites chosen are referred to as markers 2, 3 and 6, and are located at -121.3 kb, -91.4 kb and +5.6 kb (intron 1A) relative to the transcription initiation site, respectively. Only marker 2 had been characterised and deposited in public databases. Markers 3 and 6 were found by screening the human genome sequence with the Tandem Repeat Finder program at the Human Genome Browser (University of California, Santa Cruz) for regions of repetitive DNA sequence with twelve or more perfect 2-4bp repeat units close to, and then moving upstream from, the LPR. Primers were designed to generate 150-250bp fragments and PCR conditions optimised. In all, six novel putative STR loci were discovered using this technique and optimal PCR conditions were determined for five.

To be useful levels of polymorphism in the STR loci must be sufficiently high. This was determined by genotyping eight individuals from the first four Antioquian BPI family units using PCR with a fluorescently labelled forward primer and unlabelled reverse primer, and analysed on an acrylamide gel using an ABI 377 Sequencer. To be deemed sufficiently polymorphic for this study two or more

different alleles needed to be present in unrelated individuals from these eight individuals (16 chromosomes).

4.2.3 Genotyping

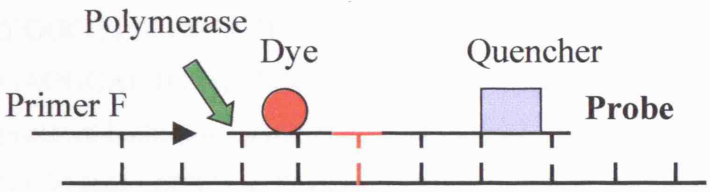
4.2.3.1 SNPs

SNPs were typed using the Taqman® assay, which is summarised in figure 4.2.1. In short, the Taqman® assay involves designing a short, approximately 20bp probe that has a FAM or VIC reporter dye at the 5' end and a nonfluorescent quencher at the 3' end. An assay for each bi-allelic SNP contains a FAM and VIC probe, one specific for each allele, as well as a primer pair specific for flanking sequence. Under reaction conditions the probe binds to the target sequence. As PCR amplification occurs, polymerase separates the probe dinucleotides thereby disrupting the association between reporter dye and quencher. The reporter dye is then free to fluoresce and can be detected. If a probe is not specific to an allele then the quencher and reporter dye will remain intact and no fluorescence will be detected.

Fluorescently labelled probes were either available immediately from the Assays-on-Demand service (markers 7 and 10), or probes were designed specifically for a target region using the Assay-by-Design facility at ABI. Probe design was unsuccessful for the second most 5' flanking SNP. The Taqman® assay involves a simultaneous probe annealing/amplification reaction in a 5µl (microlitres) reaction mixture. For the Assay-by-Demand this consists of 2.5µl Taqman® Universal PCR Master Mix, No AmpErase® UNG (2X), 0.25µl 20X Assays-on-Demand™ SNP Genotyping Assay Mix, 10ng (nanograms) genomic DNA, and made up to 5µl with distilled H_2O. A 20X Assays-on-Demand™ SNP Genotyping Assay Mix consists of 18µM (micromolar) each primer and 4µM probe. For the Assay-by-Design the reaction mixture consists of 2.5µl Taqman® Universal PCR Master Mix, No AmpErase® UNG (2X), 0.125µl 40X Assays-by-Design Assay Mix, 10ng genomic DNA, and made up to 5µl with distilled H_2O. A 40X Assays-by-Design Assay Mix consists of 36µM each primer and 8µM probe. The cycling protocol includes an initial denaturation step of 95°C for 10 mins., followed by 40 cycles of 92°C for 15 secs. (denature); 60°C for 1 min. (anneal/extend). These reactions were carried out in Perkin Elmer 9700 GeneAmp® PCR System thermocyclers.

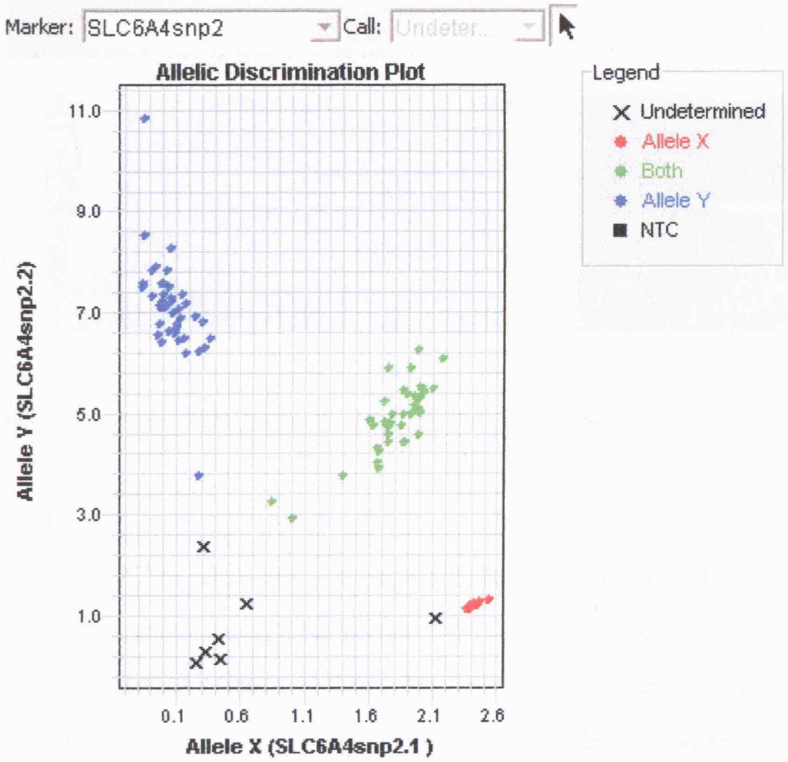
Detection of fluorescence was carried out by an ABI PRISM® 7900 HT Sequence Detection System and genotypes were read and analysed using the SDS v2.1 software. Figure 4.2.2 shows a typical output.

Figure 4.2.1. The Taqman® Assay.



The probe, with VIC or FAM dye and quencher, bonds to one of the SNP alleles. As the polymerase proceeds along the template DNA strand the bonds between the probe nucleotides are broken, the dye and quencher are dissociated and fluorescence can be detected by a 7900 HT machine.

Figure 4.2.2. Output of Taqman assay using an ABI PRISM® 7900 HT Sequence Detection System and analysed using the SDS v2.1 software.



Allele X is always the VIC labelled probe and allele Y the FAM.

4.2.3.2 LPR

The *SLC6A4* LPR was typed using PCR under the following conditions. 40ng of genomic DNA were amplified in a 20µl reaction using 5 pmol (picomoles) of each primer, 200µM dA/T/CTP, 100µM dGTP, 100µM 7-Deaza-dGTP, 1X Perkin Elmer PCR Buffer II, 1.5mM MgCl2, 2% DMSO, 3 units Amplitaq Gold and made up to 20µl using distilled H_2O. The forward primer was 5HTTLPR_F, 5'GGCGTTGCCGCTCTGAATGC, and reverse primer was 5HTTLPR_R, 5' GAGGGACTGAGCTGGACAACCAC. Thermal cycles followed a step-down protocol including an initial denaturation step of 94°C for 5 mins.; 2 cycles of 94°C for 30 secs.: 66°C for 30 secs.: 72°C for 1 min.; 3 cycles of 94°C for 30 secs.: 65°C for 30 secs.: 72°C for 1 min.; followed by 3 cycles of 94°C for 30 secs.: 64°C for 30 secs.: 72°C for 1 min.; 35 cycles of 94°C for 30 secs.: 63°C for 30 secs.: 72°C for 1 min.; and completed with a 10 min. extension step at 72°C. This was run on a Perkin Elmer 9700 GeneAmp® PCR System. The PCR products were loaded into an ethidium bromide stained, 2.5% concentration agarose gel, separated by electrophoresis in 1X TBE (tris-borate EDTA) for 1 hour 40 mins. at 100V (volts), and viewed by ultraviolet light on a UVP BioDoc-It™ System UV Transilluminator.

Figure 4.2.3. Agarose gel showing LPR genotypes.

Eight samples are loaded alongside 10µl of 100bp ladder.

4.2.3.3 STRs

Specific PCR conditions and protocols for the three STR loci are summarised in table 4.2.1. For each locus 20ng genomic DNA were amplified in a 10µl reaction consisting of 5µl of 2X Qiagen Hotstart Mastermix, 5 pmol each of forward and reverse primer and made to 10 µl using distilled H_2O. At 1X, Qiagen Hotstart MasterMix contains 0.05 units/µl of Hostart Taq polymerase, 1.5mM $MgCl_2$ and

200μM dNTPs. For markers 3 and 6, thermal cycling followed a standard protocol whereas marker 2 required a step-down cycling protocol. PCRs were run on MJ Research PTC-200 Peltier Thermal Cyclers.

PCR products were run on agarose gels during optimisation experiments. Products were loaded in 2.5% agarose gels made with ethidium bromide and separated by electrophoresis using a 100V current for 40 mins. They were then viewed using an ultraviolet light source on a UVP BioDoc-It™ System UV Transilluminator.

Once clean product was achieved a fluorescent dye was added as a 5' modification to the forward primer, and the PCR was repeated on 8 individuals from the first four Antioquian families to assess level of polymorphism. 1.2 μl of diluted PCR product (table 4.2.2) was loaded in a 20% acrylamide gel (SequaGel® XR, National Diagnostics) in 1X TBE buffer, an electrophoretic current of 3 kV passed for 3.5 hours and the fluorescent alleles detected by laser technology with an ABI 377 Sequencer. Into each well 0.3 μl blue loading buffer, 1.2 μl deionised formamide and 0.5 μl of GeneScan®-350 [TAMRA]™ were also loaded. Genescan® and Genotyper® software were used to read and analyse data. Once the 3 STRs were selected all Antioquian and Costa Rican samples were genotyped in this way, with the 3 PCR products being diluted and pooled before loading. Pooling was made possible by the fact only 2 of the 3 STRs had overlapping allele size ranges, and they each had a differently coloured fluorescent label.

Table 4.2.1. Summary of primer sequence and thermal cycle protocols for the 3 STRs used in this study.

STR locus	Forward primer (5' to 3')	PCR protocol
M 6	F: CAGGAGCGAAACTCCATCTC R: TCACAAGCAATTCATGTCACC	94°C 15m 94°C 30s, 55°C 15s, 72°C 30s X10 89°C 30s, 55°C 15s, 72°C 30s X20 72°C 10m
M 3	F: CATTATGGATTACTGAGCAATTAAC R: CCAACAATGTGATTTGATGAC	94°C 15m 94°C 30s, 55°C 15s, 72°C 30s X10 89°C 30s, 55°C 15s, 72°C 30s X20 72°C 10m
M 2	F: CAACGGTAGGCAAGTAAGCT R: TGCTGTGTGCAGTTTGATCT	94°C 15m 94°C 30s, 56°C 15s, 72°C 30s X3 94°C 30s, 55°C 15s, 72°C 30s X3 94°C 30s, 54°C 15s, 72°C 30s X3 89°C 30s, 53°C 15s, 72°C 40s X21 72°C 10m

Table 4.2.2. Size ranges and fluorescent labels of STR loci.

STR locus	Allele Size Range (bp)	Fluorescent label	Final dilution factor
M 6	182-221	HEX (yellow)	1:6
M 3	164-172	FAM (blue)	1:10
M 2	158-177	TET (green)	1:24

Colour indicated is for ABI 377 Genotyper filter set C.

4.2.4 Statistical Analyses

Genotypic data was collected and checked for Mendelian inconsistencies using the UNKNOWN program from the LINKAGE software package.

4.2.4.1 Measures of Genetic Diversity

Heterozygosity

Locus heterozygosity was measured as the total sample frequency of all heterozygotes, for all combinations of alleles, and can be calculated by summing the numbers of all heterozygotes at a locus and dividing by the sample size.

$$\text{Locus Heterozygosity} = \frac{\text{sum all heterozygotes}}{n}$$

where n is the sample size.

Average heterozygosity was calculated as the sum of all locus heterozygosities divided by the total number of loci.

$$\text{Average Heterozygosity} = \frac{\text{sum all locus heterozygosities}}{m}$$

where m is the total number of loci.

Haplotype Diversity

Haplotype diversity was calculated from Nei's gene diversity considering each haplotype as a separate allele (Nei, 1987) (see section 3.2.4.5).

4.2.4.2 Linkage Disequilibrium

LD analysis was conducted on both transmitted chromosomes and those that were not transmitted to affected offspring. Haplotypes were inferred using Simwalk2 (Sobel and Lange, 1996; Sobel et al., 2002; Sobel et al., 2001) from the family units, and separated into those transmitted and non-transmitted manually. p values from Fisher's Exact test were calculated for pairwise measurements on haplotypic data using Arlequin (Schneider et al., 2000). Lewontin's D' was also calculated for pairwise measurements and expressed pictorially using GOLD (Abecasis and Cookson, 2000).

4.2.4.3 Haplotype Block Structure

Parental and affected chromosomes were entered into the HaploView program (Barrett et al., 2005) to determine LD block structure in *SLC6A4*, for both populations separately. LD was measured by D' and block structure was defined using the 'solid spine' option. HaploView implements the expectation-maximisation algorithm to infer maximum likelihood haplotypes for each LD block, which were then analysed for association separately.

4.2.4.4 Disease Association

TDT, from the GENEHUNTER software package, was performed on individual markers and multi-marker haplotypes (Kruglyak et al., 1996). The TDT assesses whether either allele at a candidate locus is preferentially transmitted from heterozygous parents to the affected offspring. If a biallelic marker is considered this is tabulated;

	Non-Transmitted		
Transmitted	Allele1	Allele2	Total
Allele1	a	b	a + b
Allele2	c	d	c + d
Total	a + c	b + d	2n

where n = number of trios. The test statistic is calculated from

$\chi^2_{td} = (b-c)^2 / (b+c)$,

which is asymptotically chi-squared with 1 degree of freedom. It is testing the deviation from an equal distribution of b and c classes, calculated from (b + c)/2, expected under equilibrium, assuming no segregation distortion.

The HHRR test (haplotype-based haplotype relative risk) was also performed on individual markers and multi-marker haplotypes using TRANSMIT (Clayton, 1999). The HHRR performs a case-control type analysis using transmitted alleles from both heterozygous and homozygous parents as the case group and their non-transmitted counterparts as the controls. For a bi-allelic marker this can be tabulated;

	Non-Transmitted		
Transmitted	Allele1	Allele2	Total
Allele1	t_{11}	t_{12}	$t_{1.}$
Allele2	t_{21}	t_{22}	$t_{2.}$
Total	$t_{.1}$	$t_{.2}$	$t_{..}$

where '.' represents either allele 1 or 2. The test statistic is calculated by

$$HHRR = \frac{(t_{1.} - t_{.1})^2}{(t_{1.} + t_{.1})^2} + \frac{(t_{2.} - t_{.2})^2}{(t_{2.} + t_{.2})^2}$$

which is asymptotically chi-squared, under linkage equilibrium, with 1 degree of freedom (Sham, 1997).

Both these analyses were done at the Genetic Linkage User Environment (GLUE) at MRC-RFCGR (Hinxton, Cambridge). Additionally, to increase sample size and power of tests, population samples were combined and analyses repeated. Power of tests were calculated using the TDT Power Calculator program (Chen and Deng, 2001) assuming the BPAD parameters estimated by Freimer *et al* (1996).

4.2.4.5 Websites

The following websites were used in this study:

ENSEMBL: http://www.ensembl.org/index.html

The SNP Consortium: http://snp.cshl.org/

Tandem Repeat Finder and Draft Human Gene Browser: http://genome.ucsc.edu/

The LINKAGE package: http://linkage.rockefeller.edu

Simwalk2: http://watson.hgen.pitt.edu/docs/simwalk2.html

GOLD: http://www.sph.umich.edu/csg/abecasis/GOLD/

The 'TDT' option in GENEHUNTER and TRANSMIT were both used via the Genetic Linkage User Environment (GLUE) at MRC-RFCGR (Hinxton, Cambridge): http://www.hgmp.mrc.ac.uk

Haploview: http://www.broad.mit.edu/mpg/haploview/index.php

4.3 Results

4.3.1 Marker Information

4.3.1.1 SNPs and LPR

Data on the LPR and SNPs used is summarised in table 4.3.1, including rs numbers, physical locations, allele frequencies and marker heterozygosities for both populations. Calculations were done on parents only.

4.3.1.2 STRs

Information on the number of alleles, heterozygosities and locations for each STR is presented in table 4.3.2. Population frequencies are shown in table 4.3.3. marker 6 was selected after screening approximately 37.9kb around the LPR (marker 5), using the Tandem Repeat Finder function at the UCSC Genome Browser, in order to further define LPR haplotypes that may be involved in BP. Four other STRs were discovered during screening but were rejected based on poor PCR or low diversity in a sub-sample of individuals. Markers 2 and 3 were selected based on their physical location in order to augment the low amount of information supplied by the low minor allele frequency 5' SNPs. To my knowledge neither marker 3 nor marker 6 have been characterised before. Marker 2 (D17S1532), on the other hand, is a known microsatellite and is stored in the UniSTS database at NCBI, although no heterozygosity data was available. In most cases STRs performed well. One notable exception is for marker 3 which did not amplify well in Antioquia samples, deviated significantly from Hardy-Weinberg equilibrium ($p=0.0003$) in this population and was subsequently removed from LD analysis.

Figure 4.3.1 shows the positions of all markers relative to *SLC6A4*.

Table 4.3.1. Summary of bi-allelic markers used.

Marker	Rs Number	Map Location[a]	Allele	Allele Frequency		H		HWE p value	
				Ant n 96	CVCR n 107	Ant	CVCR	Ant	CVCR
1	887469	28,838,137	1=c 2=t	0.9167 0.0833	0.9563 0.0437	0.1667	0.0874	1.0000	1.0000
4	2129785	28,736,093	1=a 2=g	0.9421 0.0579	0.919 0.081	0.0947	0.1619	0.2531	1.0000
5	5HTT-LPR	28,709,687 to 28,710,015[b]	1=S 2=L	0.4628 0.5372	0.545 0.455	0.5426	0.5100	0.4203	0.8453
7	2066713	28,697,228	1=a 2=g	0.3681 0.6319	0.3465 0.6535	0.5165	0.5149	0.3584	0.2603
8	2020939	28,696,295	1=t 2=c	0.4086 0.5914	0.5049 0.4951	0.5376	0.5588	0.4003	0.3184
9	2020942	28,692,477	1=g 2=a	0.5989 0.4011	0.6535 0.3465	0.5385	0.5149	0.2947	0.2700
10	140701	28,684,095	1=c 2=t	0.6146 0.3854	0.4806 0.5194	0.5625	0.5340	0.0875	0.5553
11	3813034	28,670,537	1=t 2=g	0.5895 0.4105	0.4764 0.5236	0.5895	0.5189	0.0584	0.8469
12	4494608	28,634,354	1=g 2=a	0.1368 0.8632	0.1068 0.8932	0.2526	0.1942	1.0000	1.0000
13	4310926	28,578,838	1=a 2=g	0.5598 0.4402	0.4563 0.5437	0.5761	0.5243	0.1369	0.6884
14	4239227	28,534,666	1=t 2=g	0.4086 0.5914	0.5194 0.4806	0.5806	0.5340	0.0550	0.5563
					Mean	0.4507	0.4230		

[a]Physical coordinates according to the draft human genome browser (July 2004). [b]Size given as long allele. H: heterozygosity; HWE: Hardy-Weinberg Equilibrium; Ant: Antioquia; CVCR: Central Valley of Costa Rica.

Table 4.3.2. Summary of STRs used.

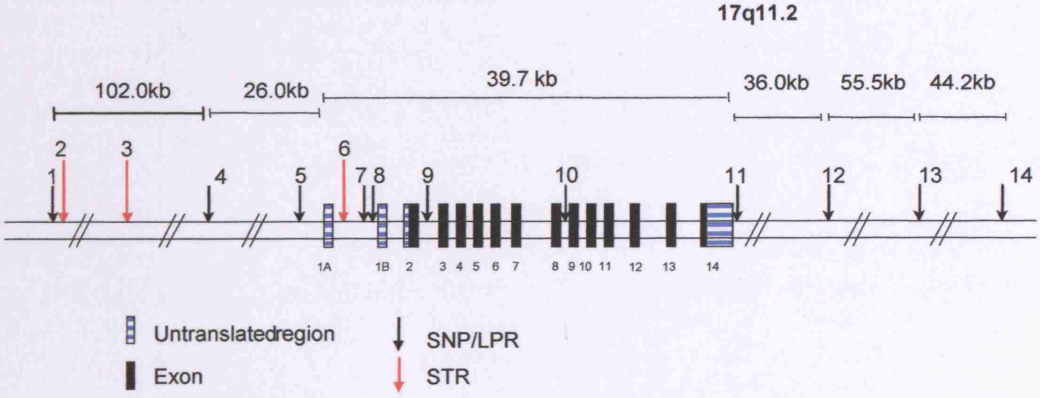
Marker	Map Location[a]	Repeat Unit	Antioquia			CVCR		
			No. Alleles	H	HWE p value	No. Alleles	H	HWE p value
2	28,829,642	ATT	7 (n 95)	0.6105	0.2106	7 (n 88)	0.5114	0.0622
3	28,799,627	AC	6 (n 58)	0.2586	0.0003	6 (n 87)	0.4253	0.7241
6	28,702,604	AAG	17 (n 89)	0.8652	0.3369	14 (n 93)	0.9140	0.7316
			Mean	0.5781		Mean	0.6169	

[a]Positions of first basepair in first repeat unit according to UCSC draft genome browser (July 2004). H: heterozygosity, HWE: Hardy-Weinberg Equilibrium.

Table 4.3.3. STR allele frequencies in Antioquia and CVCR.

Marker 2			Marker 3			Marker 6		
Allele	Freq		Allele	Freq		Allele	Freq	
	Ant	CVCR		Ant	CVCR		Ant	CVCR
1	0.0000	0.0170	1	0.0086	0.0057	1	0.0056	0.0000
2	0.0053	0.0000	2	0.0086	0.0057	2	0.0056	0.0000
3	0.0526	0.0682	3	0.7500	0.7069	3	0.0056	0.0000
4	0.0263	0.0227	4	0.1121	0.2011	4	0.0056	0.0054
5	0.5526	0.6420	5	0.0948	0.0747	5	0.0674	0.0430
6	0.2842	0.1648	6	0.0259	0.0000	6	0.0337	0.0215
7	0.0737	0.0795	7	0.0000	0.0057	7	0.0899	0.0269
8	0.0000	0.0057				8	0.1067	0.0699
9	0.0053	0.0000				9	0.1124	0.0860
						10	0.0955	0.1559
						11	0.2022	0.2581
						12	0.1629	0.1183
						13	0.0674	0.1452
						14	0.0169	0.0430
						15	0.0056	0.0054
						16	0.0112	0.0054
						17	0.0000	0.0161
						18	0.0056	0.0000
No. Alleles	7	7		6	6		17	14

Figure 4.3.1. The *SLC6A4* gene with positions of markers.



4.3.1.3 Non-transmitted and Transmitted Chromosomes

As my DNA sample includes DNA from BPI trios, it is important to analyse those chromosomes transmitted to affected offspring and those non-transmitted separately. Allele frequencies, marker diversities as well as eleven marker haplotype frequencies and diversities are shown in tables 4.3.4 to 4.3.7.

Table 4.3.4. Allele frequencies for the bi-allelic markers in chromosomes not transmitted and transmitted to the affected offspring.

		Antioquia		CVCR	
Marker	**Allele**	**NT**	**Trans**	**NT**	**Trans**
1	1	0.9468	0.9167	0.9697	0.9286
	2	0.0532	0.0833	0.0303	0.0714
4	1	0.9355	0.9580	0.9596	0.9286
	2	0.0645	0.0420	0.0404	0.0714
5	1	0.4000	0.4965	0.6042	0.5364
	2	0.6000	0.5035	0.3958	0.4636
7	1	0.3596	0.3379	0.3196	0.3247
	2	0.6404	0.6621	0.6804	0.6753
8	1	0.3516	0.4406	0.5556	0.5065
	2	0.6484	0.5594	0.4444	0.4935
9	1	0.6067	0.6345	0.6907	0.6688
	2	0.3933	0.3655	0.3093	0.3312
10	1	0.7021	0.5694	0.4444	0.4837
	2	0.2979	0.4306	0.5556	0.5163
11	1	0.6667	0.5625	0.4242	0.4740
	2	0.3333	0.4375	0.5758	0.5260
12	1	0.1398	0.1250	0.1546	0.0921
	2	0.8602	0.8750	0.8454	0.9079
13	1	0.6667	0.5177	0.3878	0.4610
	2	0.3333	0.4823	0.6122	0.5390
14	1	0.3187	0.4375	0.5612	0.5195
	2	0.6813	0.5625	0.4388	0.4805

NT: non-transmitted, Trans: transmitted.

Table 4.3.5. Allele frequencies for the STR markers in chromosomes not transmitted and transmitted to the affected offspring.

		Antioquia		CVCR	
Marker	**Allele**	**NT**	**Trans**	**NT**	**Trans**
2	1	0.000	0.000	0.012	0.007
	2	0.011	0.007	0.000	0.000
	3	0.065	0.042	0.081	0.073
	4	0.032	0.028	0.023	0.013
	5	0.570	0.563	0.586	0.669
	6	0.247	0.278	0.161	0.185
	7	0.075	0.076	0.126	0.046
	8	0.000	0.000	0.012	0.007
	9	0.000	0.007	0.000	0.000
Marker	**Allele**	**NT**	**Trans**	**NT**	**Trans**
3	1	0.018	0.000	0.000	0.007
	2	0.018	0.031	0.012	0.000
	3	0.696	0.763	0.647	0.743
	4	0.125	0.134	0.247	0.167
	5	0.107	0.062	0.094	0.076
	6	0.036	0.010	0.000	0.000
	7	0.000	0.000	0.000	0.007
Marker	**Allele**	**NT**	**Trans**	**NT**	**Trans**
6	1	0.012	0.000	0.000	0.000
	2	0.000	0.007	0.000	0.000
	3	0.000	0.007	0.000	0.000
	4	0.000	0.007	0.000	0.007
	5	0.081	0.043	0.044	0.039
	6	0.058	0.021	0.000	0.039
	7	0.081	0.085	0.055	0.013
	8	0.115	0.085	0.088	0.072
	9	0.058	0.135	0.055	0.118
	10	0.092	0.135	0.165	0.157
	11	0.195	0.206	0.242	0.268
	12	0.172	0.163	0.143	0.118
	13	0.092	0.043	0.143	0.098
	14	0.012	0.057	0.044	0.033
	15	0.012	0.000	0.000	0.013
	16	0.012	0.007	0.000	0.007
	17	0.000	0.000	0.022	0.020
	18	0.012	0.000	0.000	0.000

NT: non-transmitted, Trans: transmitted.

Table 4.3.6. Haplotype frequencies and diversities in non-transmitted and transmitted chromosomes in Antioquia and CVCR, generated from all bi-allelic markers.

5' to 3'	Ant NT	Ant Trans	CVCR NT	CVCR Trans
11212211212	0.156	0.137	0.159	0.127
11121122221	0.117	0.214	0.352	0.327
11221122221	0.091	0.053	0.034	0.040
11222111112	0.078	0.031	0.023	0.020
11112211212	0.078	0.061	0.023	0.053
11121211212	0.052	0.000	0.000	0.000
12221122221	0.039	0.000	0.023	0.027
11222111222	0.039	0.000	0.034	0.000
21122111112	0.026	0.023	0.023	0.000
11122122221	0.026	0.000	0.000	0.000
11122111212	0.026	0.000	0.000	0.027
11122111222	0.026	0.000	0.000	0.000
11222111212	0.026	0.031	0.000	0.020
21212211212	0.000	0.031	0.000	0.000
11212111212	0.000	0.023	0.000	0.000
11222211212	0.000	0.000	0.023	0.000
11122211112	0.000	0.000	0.023	0.000
11122111112	0.000	0.000	0.023	0.033
11212211112	0.000	0.000	0.000	0.020
rare haps	0.221	0.397	0.261	0.307
Div	0.895	0.773	0.785	0.780
+/- variance	+/- 0.017	+/- 0.023	+/- 0.027	+/- 0.021

STRs were not included in these haplotypes to limit the number of rare haplotypes. NT: non-transmitted, Trans: transmitted, rare haps: haplotypes less than 2%, Div: haplotype diversity considering each haplotype (and the collective rare haplotypes) as a separate allele.

Table 4.3.7. π (equivalent here to average gene diversity over all marker loci) for non-transmitted and transmitted chromosomes in both the Antioquia and CVCR populations (+/-variance).

	Antioquia	CVCR
NonTransmitted	0.4392 +/- 0.2360	0.4366 +/- 0.2347
Transmitted	0.4483 +/- 0.2403	0.4636 +/- 0.2477

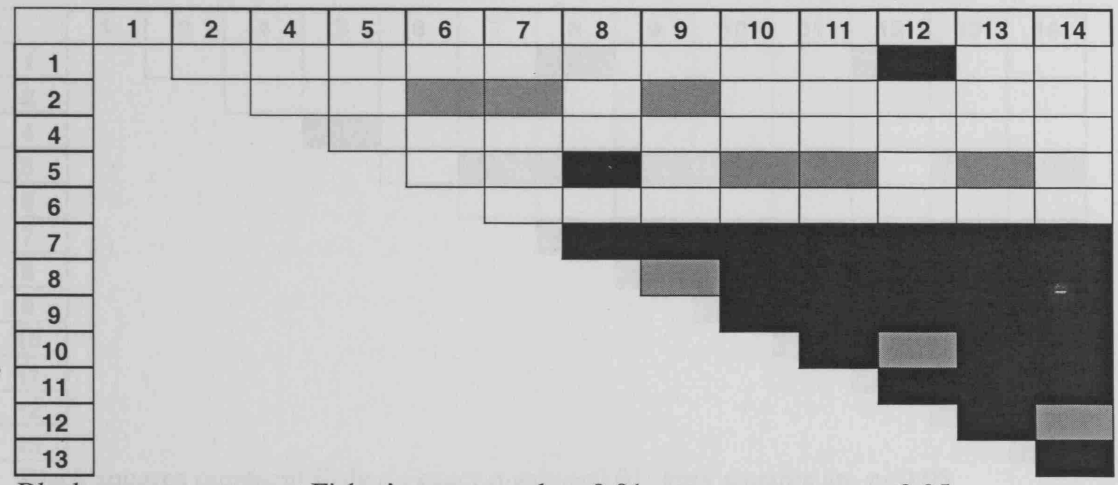
4.3.2 Linkage Disequlibrium

4.3.2.1 FETp

Figures 4.3.2 and 4.3.4 show significant LD in non-transmitted chromosomes, measured by Fisher's exact test, and reveal that the level of background linkage disequilibrium in both the Antioquians and Costa Ricans is roughly similar. In general, an LD block begins from marker 7 and extends at least as far as marker 14, which spans a distance of 162,562 kb. In both populations the LPR (marker 5) appears to maintain some level of LD with the coding part of the gene. One notable difference between the two populations is that the 3' block does not seem to be as strong in CVCR as marker 13 is not in LD with markers 7 to 9.

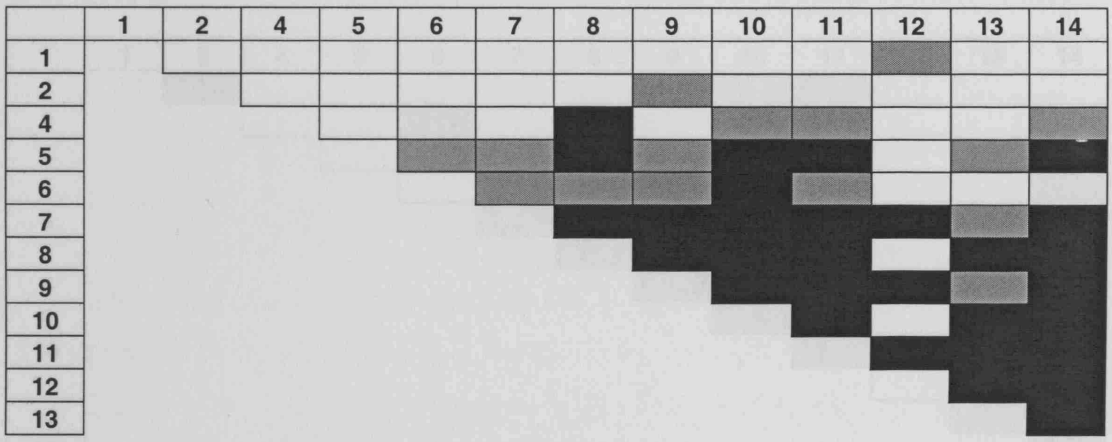
Interestingly, there seems to be some discrepancy in LD pattern in the 5' region of the *SLC6A4* gene in transmitted versus non-transmitted chromosomes, and this trend is repeated in both populations. By comparing figures 4.3.3 and 4.3.5 to 4.3.2 and 4.3.4, respectively, the 5' region of the LD block includes marker 6 in the transmitted chromosomes, whereas it is not included in their non-transmitted counterparts. A decrease in genetic diversity in the transmitted chromosomes may also be expected if this were the case, though this was not observed (tables 4.3.6 and 4.3.7); although a lower diversity is seen in the Antioquian transmitted chromosomes, this effect is likely exaggerated by the large number of pooled rare haplotypes here.

Figure 4.3.2. Pairwise LD in Antioquia non-transmitted chromosomes.



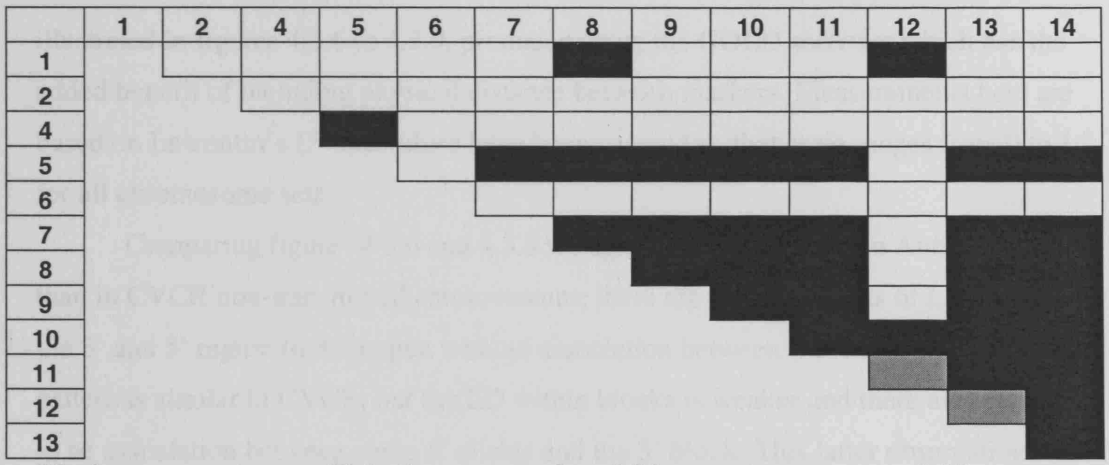
Black squares represent Fisher's exact p value<0.01, grey squares are p<0.05.

Figure 4.3.3. Pairwise LD in Antioquia transmitted chromosomes.



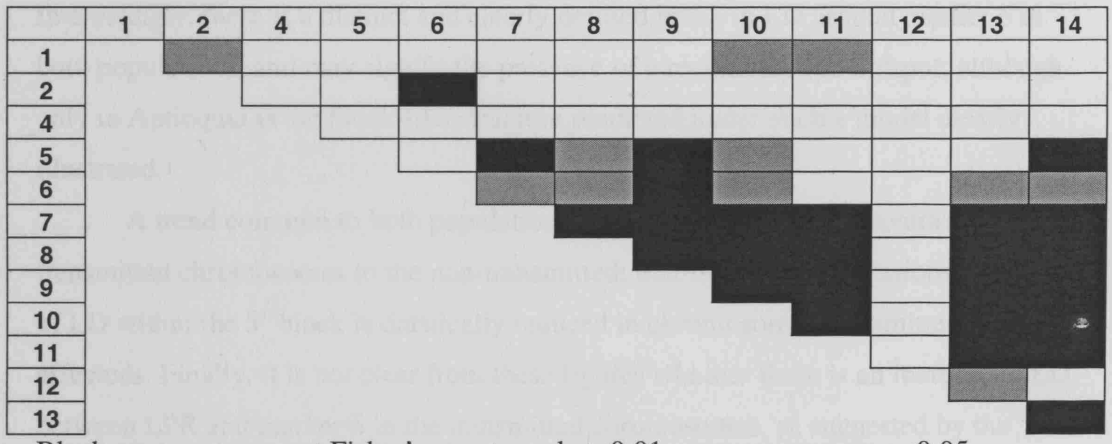
Black squares represent Fisher's exact p value<0.01, grey squares are p<0.05.

Figure 4.3.4. Pairwise LD in CVCR non-transmitted chromosomes.



Black squares represent Fisher's exact p value<0.01, grey squares are p<0.05.

Figure 4.3.5. Pairwise LD in CVCR transmitted chromosomes.



Black squares represent Fisher's exact p value<0.01, grey squares are p<0.05.

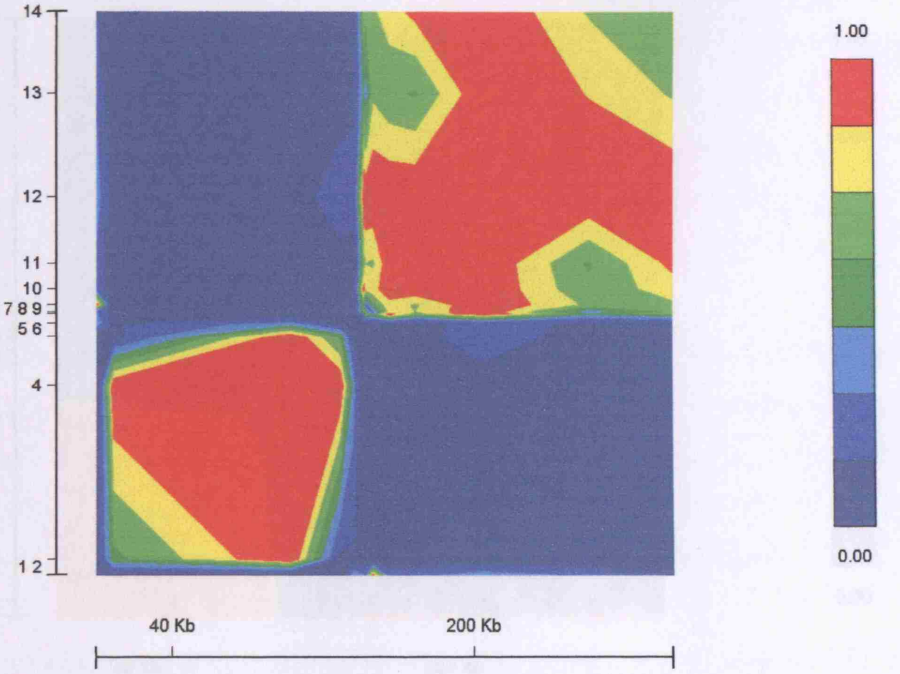
4.3.2.2 GOLD results

Patterns of LD in the four different classes of chromosomes are further illustrated in figures 4.3.6 to 4.3.9, produced using the GOLD software which has the added benefit of including physical distance between markers. Measurements here are based on Lewontin's D' and values have been squared so that scale ranges from 0 to 1 for all chromosome sets.

Comparing figures 4.3.6 and 4.3.8 we again see stronger LD in Antioquian than in CVCR non-transmitted chromosomes; there are distinct blocks of LD both in the 3' and 5' region in Antioquia with no association between these two blocks. This pattern is similar in CVCR, but the LD within blocks is weaker and there also appears to be association between some 5' alleles and the 3' block. This latter observation may be an artefact of the very low minor allele frequencies of markers 1 and 4 in CVCR, as low minor allele frequencies are known to inflate values of D' (Ardlie et al., 2002). From marker 14, strong LD extends as far as marker 7 in both populations, supporting the FETp values. However, in contrast to findings based on FETp values, the LPR is not shown to be in LD with the coding region of the gene here whatsoever. Interestingly, there is a distinct and clearly defined break in LD around marker 6 in both populations, and may signify the presence of a recombination hotspot; although only in Antioquia is the block-like structure predicted under such a model clearly illustrated.

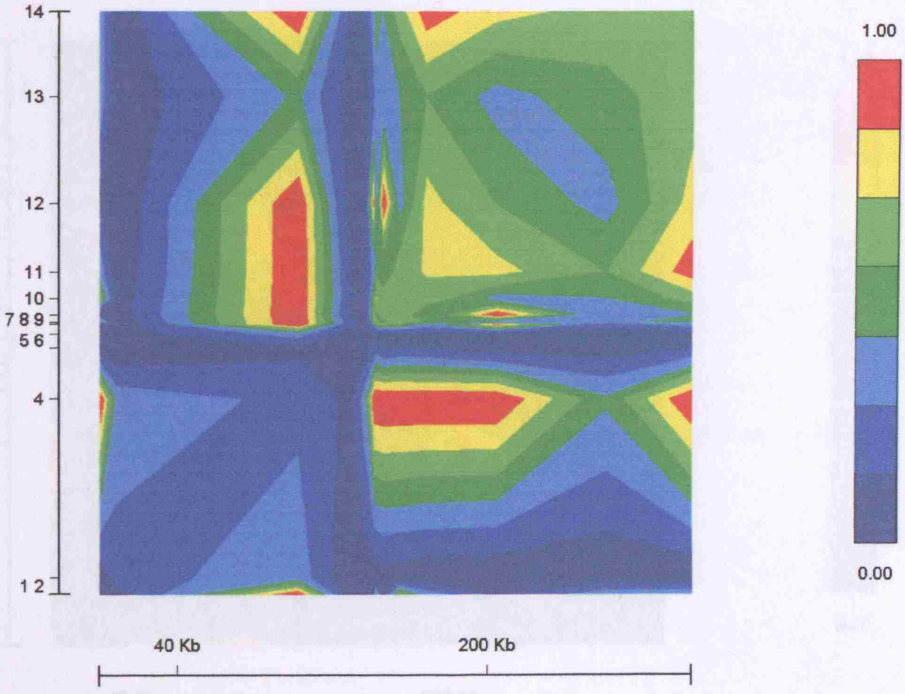
A trend common to both populations is made apparent when contrasting the transmitted chromosomes to the non-transmitted; that is in both populations the level of LD within the 5' block is drastically reduced in chromosomes transmitted to the affecteds. Finally, it is not clear from these figures whether there is an increase in LD between LPR and marker 6 in the transmitted chromosomes, as suggested by the FETp values.

Figure 4.3.6. GOLD figure of LD in non-transmitted chromosomes in Antioquia.



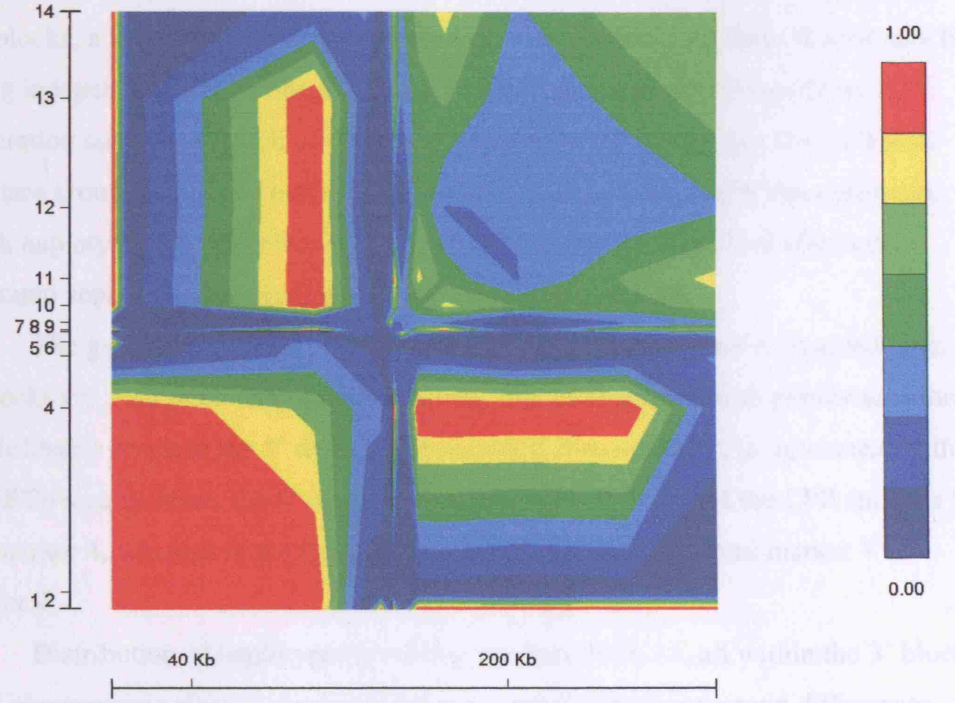
A value of 1.00 represents complete LD as measured by Lewontin's D'.

Figure 4.3.7. GOLD figure of LD in transmitted chromosomes in Antioquia.



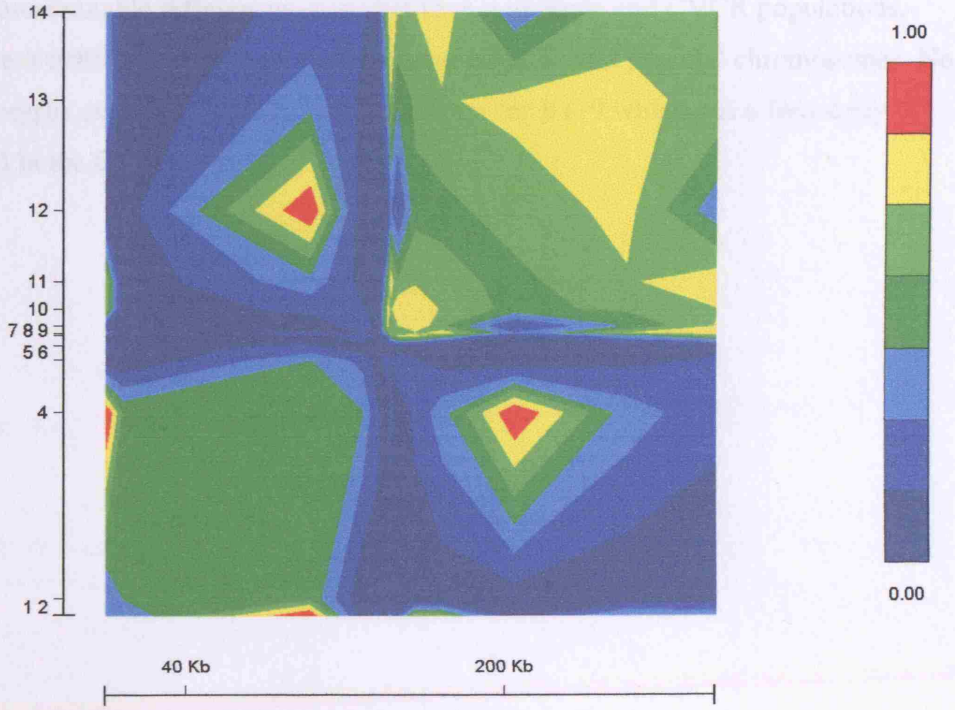
A value of 1.00 represents complete LD as measured by Lewontin's D'.

Figure 4.3.8. GOLD figure of LD in non-transmitted chromosomes in CVCR.



A value of 1.00 represents complete LD as measured by Lewontin's D'.

Figure 4.3.9. GOLD figure of LD in transmitted chromosomes in CVCR.



A value of 1.00 represents complete LD as measured by Lewontin's D'.

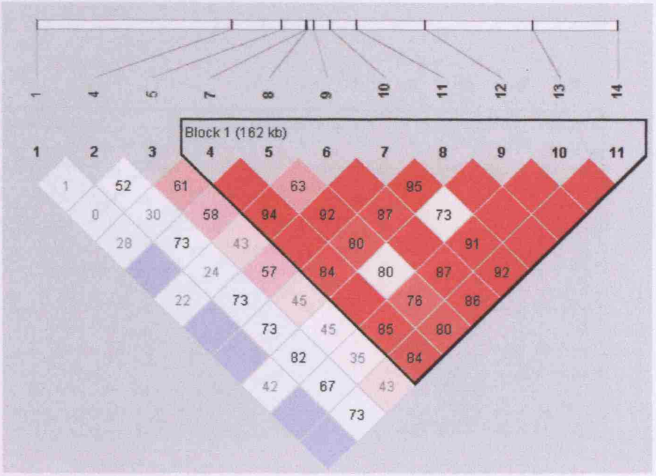
4.3.2.3 Haplotype Block Structure

The Arlequin and GOLD results show the LD in *SLC6A4* to be structured into two blocks, a large 3' block and a less well defined 5' block. As these blocks may be acting independently from one another, it is important to perform population and association analyses on each of the blocks separately. To investigate the LD block structure around *SLC6A4* I entered the bi-allelic data into the HaploView program. Block haplotypes and their frequencies in parental chromosomes and affecteds, generated separately, are presented in figures 4.3.10 to 4.3.13.

The general trend of strong LD in the 3' region of the gene is repeated here, as 3' blocks are seen in all chromosome classes. Interestingly, for both populations there is a definable block in the 5' region in transmitted chromosomes, in agreement with the FETp results. Here, the 5' block in Antioquians only includes the LPR (marker 5) and marker 4, whereas in the Costa Ricans the block stretches from marker 7 to marker 4.

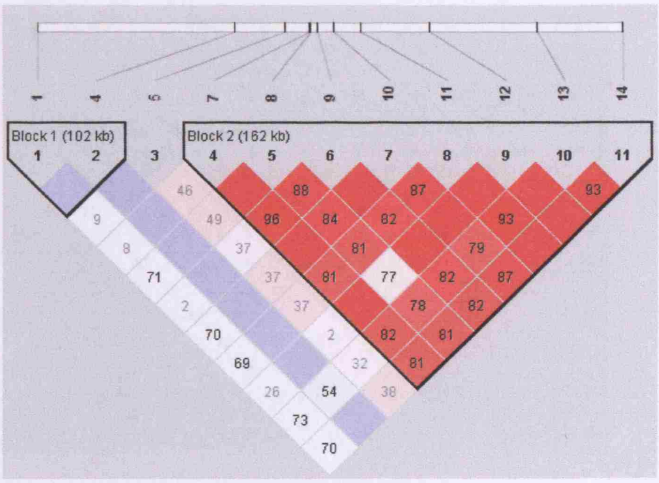
Distribution of haplotypes involving markers 10 to 14, all within the 3' block in all chromosome classes, was assessed to investigate between group differences (table 4.3.8 and figure 4.3.14). In general, three haplotypes predominate in all groups and account from 89.6% (Antioquia cases) to 96.2% (CVCR cases) of all haplotypes. The most notable differences occur between Antioquia and CVCR populations, whereas little differences are seen between affected and parental chromosomes. No haplotypes occur exclusively in cases, except for B1_9 which has a frequency of 0.013 in the CVCR cases.

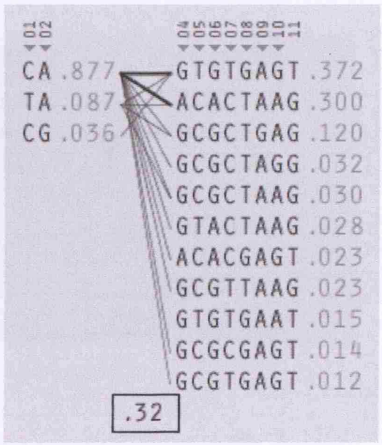
Figure 4.3.10. The LD block structure in the Antioquia parental chromosomes and the constituent haplotypes.



Markers included in the haplotypes correspond to the highlighted region of the LD figure; the numbers to the side of each haplotype are the frequencies.

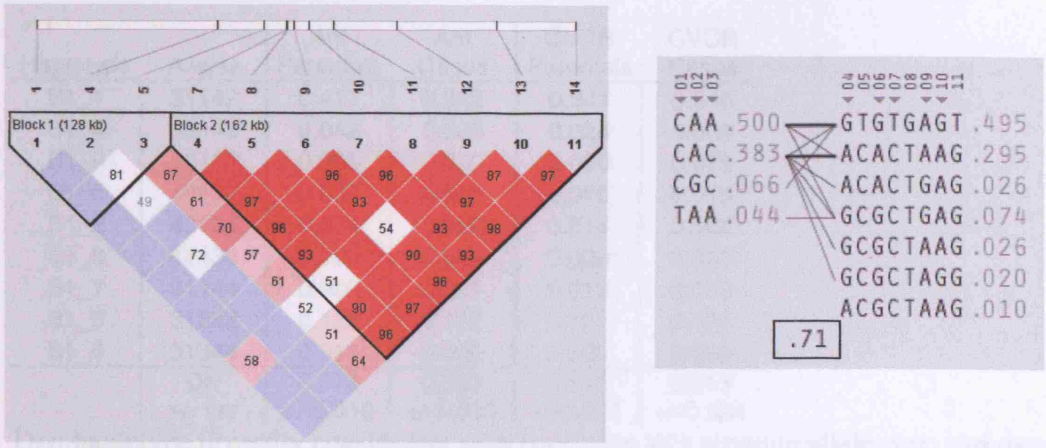
Figure 4.3.11. The LD block structure in the Antioquia case chromosomes and the constituent haplotypes.





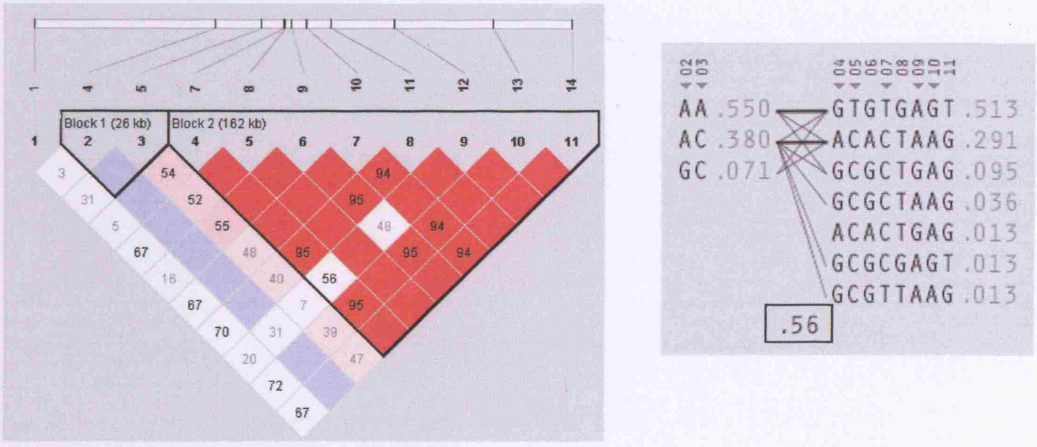
Markers included in the haplotypes correspond to the highlighted region of the LD figure; the numbers to the side of each haplotype are the frequencies. The rectangled number is the strength of LD between the two blocks, measured by D'.

Figure 4.3.12. The LD block structure in the Costa Rican parental chromosomes and the constituent haplotypes.



Markers included in the haplotypes correspond to the highlighted region of the LD figure; the numbers to the side of each haplotype are the frequencies. The rectangled number is the strength of LD between the two blocks, measured by D'.

Figure 4.3.13. The LD block structure in the Costa Rican case chromosomes and the constituent haplotypes.



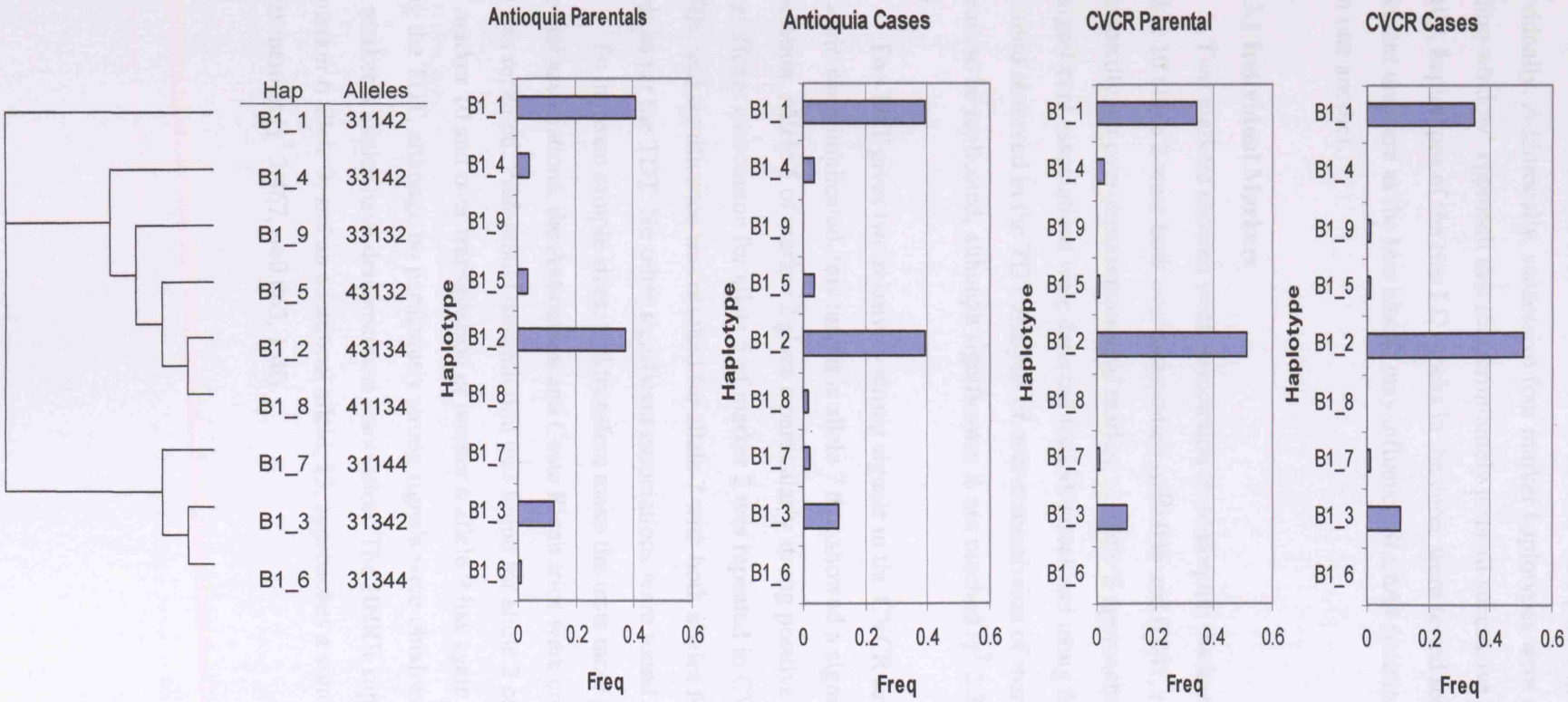
Markers included in the haplotypes correspond to the highlighted region of the LD figure; the numbers to the side of each haplotype are the frequencies. The rectangled number is the strength of LD between the two blocks, measured by D'.

Table 4.3.8. Frequencies and diversities of ML haplotypes generated from marker 10 to 14 in the 3' LD block.

Haplotype	Alleles	Ant Parentals	Ant Cases	CVCR Parentals	CVCR Cases
B1_1	31142	0.411	0.392	0.341	0.346
B1_4	33142	0.042	0.035	0.024	0.000
B1_9	33132	0.000	0.000	0.000	0.013
B1_5	43132	0.037	0.035	0.010	0.013
B1_2	43134	0.374	0.392	0.514	0.509
B1_8	41134	0.000	0.014	0.000	0.000
B1_7	31144	0.000	0.021	0.010	0.013
B1_3	31342	0.126	0.112	0.101	0.107
B1_6	31344	0.011	0.000	0.000	0.000
	Div	0.676	0.683	0.611	0.613
	+/- var	+/-0.019	+/-0.023	+/-0.021	+/-0.024

Div: haplotype diversity considering each haplotype as a separate allele; var: variance.

Figure 4.3.14. Phylogenetic relationship and distribution of *SLC6A4* 3' LD block haplotypes.



4.3.3 Association Analysis

The TDT and HHRR family association tests were performed on all markers individually. Additionally, successive four marker haplotypes were also analysed in a 'sliding window' approach that indiscriminately tests all successive haplotypes. Finally, haplotypes of the two LD blocks in the cases were tested separately; an important measure as the two blocks may influence *SLC6A4* function independently from one another.

4.3.3.1 Individual Markers

Two markers showed weak association in Antioquia; marker 6 allele 9 and marker 10 allele 2 were both over-transmitted (p=0.046 and 0.041, respectively). Additionally, an over-representation of marker 11 allele 2 approached significance. No significant associations were detected for individual loci using the HHRR test. The trend observed in the TDT analysis of over-transmission of marker 6 allele 9 appears to be replicated, although significance is not reached (χ^2 2.266, p=0.1322, 1df).

The TDT gives two relatively strong signals in the CVCR sample. Marker 6 has again been implicated, this time it is allele 7 that showed a significant negative association. Allele 5 of marker 2 gives a particularly strong positive signal (p =0.011). A significant association for allele 5 of marker 2 was repeated in CVCR using the HHRR, and significance was reached for allele 7 with both alleles following the same trends as for the TDT. No other significant associations were found.

To increase sample sizes, and therefore make the tests more powerful to detect potential associations, the Antioquian and Costa Rican trios were combined and analyses repeated. Preferential transmission was found for allele 2 of marker 12, allele 2 of marker 10 and over-transmission of marker 6 allele 9 has again been repeated using the TDT, although no particularly strong signals were obtained. Additionally, two marker 3 alleles have demonstrated association. The HHRR supports the findings for marker 6 allele 9, and an additional allele, 13, approaches a significant under-representation (χ^2 3.467, p=0.063, 1 df).

Table 4.3.9. Summary of positive and near positive association results for individual markers.

	TDT					HHRR			
	Marker / allele(s)	T	χ^2	p		**Marker / allele(s)**	T	χ^2	p
Ant	**6 /**								
	9	+	4.00	0.046					
	10 /								
	2	+	4.17	0.041					
	11 /								
	2	+	3.52	0.061					
CVCR	**2 /**					**2 /**			
	5	+	6.55	0.011		5	+	3.854	0.044
	7	-	3.57	0.059		7	-	5.204	0.021
	6 /					**6 /**			
	7	-	5.00	0.025		6	+	3.504	0.062
						7	-	3.453	0.061
Both	**3 /**					**6 /**			
	3	+	4.76	0.029		9	+	4.061	0.044
	5	-	4.50	0.034		13	-	3.467	0.063
	6 /								
	9	+	4.17	0.041					
	10 /								
	2	+	4.25	0.039					
	11 /								
	2	+	3.77	0.052					
	12 /								
	2	+	4.00	0.046					

T: Transmission (over, +, or under, -.); p: probability value; TDT: transmission disequilibrium test; HHRR: haplotype-based haplotype relative risk.

4.3.3.2 LPR

Special attention has been paid to the *SLC6A4* LPR due to its proven functionality. Analysis of the initial Antioquia dataset generated the results presented in tables 4.3.10 and 4.3.11. From these initial results no excess transmission of either allele is demonstrated. As the TDT test was only able to use information from 10 offspring a further 96 families were added to the analysis; power of the TDT analysis was increased from 0.187 to 0.52 at the 5% significance level. This gave the results presented in tables 4.3.12 and 4.3.13. An over-representation of the short allele is consistently apparent although significance is not reached. A role of either of the LPR alleles in the CVCR population is not supported here, summarised in tables 4.3.14 and 4.3.15. Finally, CVCR and the total Antioquia population were combined, which

Table 4.3.10. TDT for the LPR in the original Antioquia sample.

	trans	untrans
Allele 1 (S)	11	9
Allele 2 (L)	9	11

χ^2 0.20, p=0.655, 1 df

Table 4.3.11. HHRR for the LPR in the original Antioquia sample.

Haplotype	Observed (n 68)	Expected
1 (S)	66	64.5
2 (L)	70	71.5

χ^2 0.169, p=0.681, 1 df

n is the number of families used in the test.

Table 4.3.12. TDT for the LPR in the enlarged Antioquia sample (169 families).

	trans	untrans
Allele 1 (S)	30	23
Allele 2 (L)	23	30

χ^2 0.920, p=0.366, 1 df,

Table 4.3.13. HHRR for the LPR in the enlarged Antioquia sample.

Haplotype	Observed (n 156)	Expected
1 (S)	156	150.6
2 (L)	156	161.4

χ^2 0.978, p=0.323, 1 df

n is the number of families used in the test.

Table 4.3.14. TDT for the LPR in the CVCR.

	trans	untrans
Allele 1 (S)	12	11
Allele 2 (L)	11	12

χ^2 0.040, p=0.842, 1 df

Table 4.3.15. HHRR for the LPR in the CVCR.

Haplotype	Observed (n 76)	Expected
1 (S)	84	83.8
2 (L)	68	68.2

χ^2 0.004, p=0.950, 1 df

n is the number of families used in the test.

Table 4.3.16. TDT for the LPR in the combined sample.

	trans	untrans
Allele 1 (S)	42	34
Allele 2 (L)	34	42

χ^2 1.684, p=0.194, 1 df

Table 4.3.17. HHRR for the LPR in the combined sample.

Haplotype	Observed (n 232)	Expected
1 (S)	240	234.4
2 (L)	224	229.6

χ^2 0.1352, p=0.713, 1 df

n is the number of families used in the test.

raised the power of the TDT analysis (probability of rejecting the null hypothesis if false) at the 5% level to 0.593. Although more short alleles are transmitted than non-transmitted, significance was not reached (tables 4.3.16 and 4.3.17).

The results for the LPR show no significant allelic association of the LPR with BP. While significance has not been reached in any of the tests, there is a slight over-representation of the short allele in transmitted chromosomes, shown most strongly in the TDT test for Antioquia and CVCR combined. However, relevant to this study, recent work on the LPR has suggested that both the short and long alleles consist of sub-classes substantially increasing allelic heterogeneity at this locus. As my genotyping assay is only able to distinguish between the L (16 repeat units) and S (14 repeats) alleles, my association tests are not able to detect if one of these newly discovered alleles is associated with BP alone. If an allelic association exists the signal will be diluted by the other non-associated long or short alleles.

To address this, I have performed the same analyses using only the LPR and its adjacent heterozygous STR, marker 6. Haplotype frequencies and strength of LD between the two markers in transmitted and non-transmitted chromosomes are shown in table 4.3.18, and results from association analyses in 4.3.19. For both populations, FETp falls below 0.05 in the transmitted chromosomes only. Notably, a significant over-representation of haplotype 2 9 is demonstrated in six instances. The strongest p value is 0.0047 (TDT $\chi^2 = 8.00$) for the combined Antioquia and CVCR sample with dhskip on, which was transmitted eight times, compared to zero non-transmissions, and is actually the strongest p value attained in all association analyses. Although this measurement was based on just eight families, significance is also reached with the HHRR which estimated this haplotype was transmitted twenty-three times. Interestingly, LPR allele 2 corresponds to the long allele.

Table 4.3.18. LPR/marker 6 haplotype frequencies in transmitted and non-transmitted chromosomes, and statistical allelic association as measured by FET p values.

	Antioquia		CVCR	
Haplotype	NT	Trans	NT	Trans
1 3	0.000	0.007	0.000	0.000
1 5	0.024	0.029	0.011	0.007
1 6	0.000	0.000	0.000	0.013
1 7	0.000	0.007	0.000	0.000
1 8	0.036	0.022	0.057	0.027
1 9	0.036	0.066	0.046	0.040
1 10	0.024	0.066	0.102	0.093
1 11	0.083	0.109	0.159	0.167
1 12	0.107	0.117	0.102	0.073
1 13	0.036	0.022	0.091	0.053
1 14	0.012	0.037	0.034	0.027
1 15	0.000	0.000	0.000	0.013
1 16	0.012	0.007	0.000	0.000
1 17	0.000	0.000	0.023	0.020
1 18	0.012	0.000	0.000	0.000
2 1	0.012	0.000	0.000	0.000
2 2	0.000	0.007	0.000	0.000
2 4	0.000	0.007	0.000	0.007
2 5	0.060	0.015	0.034	0.033
2 6	0.060	0.015	0.000	0.027
2 7	0.083	0.073	0.057	0.013
2 8	0.060	0.066	0.034	0.047
2 9	0.024	0.073	0.011	0.080
2 10	0.071	0.073	0.068	0.067
2 11	0.107	0.095	0.091	0.093
2 12	0.071	0.051	0.034	0.040
2 13	0.060	0.022	0.034	0.047
2 14	0.000	0.015	0.011	0.007
2 15	0.012	0.000	0.000	0.000
2 16	0.000	0.000	0.000	0.007
FETp	0.079	0.022	0.139	0.047

Table 4.3.19. Summary of significant association results for haplotypes of LPR and marker 6 loci, for Antioquia, CVCR and the populations combined.

		TDT					HHRR			
		LPR M6	T	χ^2	p (dh skip)		LPR M6	T	χ^2	p
Ant	Alleles	2 9	+	3.57 6.00	0.059 (off) 0.014 (on)					
CVCR	Alleles	2 7	-	5.00 5.00	0.025 (off) 0.025 (on)		2 7 *	-	3.559	0.059
	Alleles	2 13	+	5.00 5.00	0.025 (off) 0.025 (on)		2 9	+	5.012	0.025
Both	Alleles	2 9	+	5.44 8.00	0.020 (off) 0.005 (on)		2 9	+	5.018	0.025

* p values based on < 5 observations. M6: marker 6; T: Transmission (over, +, or under, -.); p: probability value; TDT: transmission disequilibrium test; HHRR: haplotype-based haplotype relative risk.

4.3.3.3 Four marker 'sliding window'

Two successive four marker haplotypes demonstrated linkage in Antioquia using the TDT, involving markers 10, 11, 12, 13 and 14. Analysis was done with dhskip (a correction measure in 'TDT' for reducing type 1 error rate with multilocus haplotypes) initially turned off to ascertain potentially important loci. The HHRR produced similar non-significant trends for these haplotypes.

One four marker haplotype in the 5' region demonstrated significant association in the CVCR using the TDT, and this association is maintained at the same level with dhskip on. Interestingly, this haplotype includes allele 7 of marker 2 which showed a near significant under-representation individually. A significant association was demonstrated for nine haplotypes in the CVCR population with the HHRR, although six of these p values were based on less than five transmissions. The remaining three haplotypes are contained in two successive 5' regions involving markers 1, 2, 3, 4 and 5 (the LPR), further implicating this region. Marker 2 allele 5 is included in the two most common haplotypes giving significant positive associations, with 45 and 63 transmissions, and one of these haplotypes contains the short LPR allele. Furthermore, a global tests of association for haplotypes of markers 1, 2, 3 and 4 generated by the HHRR, produced a p value of 0.037 (12 df); and for haplotypes of 3' markers 11, 12, 13 and 14 a p value of 0.051 was produced, before correcting for multiple testing.

In the combined sample only one haplotype - involving markers 10, 11, 12 and 13 - with an adequate number of observations produced a strong signal at p=0.013 using the TDT. With dhskip on significance was almost reached at p=0.059. A relatively large number of four locus haplotypes produced significant signals under the HHRR method. These were localised to haplotypes in the 3' region involving markers 10, 11, 12, 13 and 14 as well as 5' haplotypes involving 1, 2, 3, 4 and the LPR. These results largely extend previous findings for both Antioquia and CVCR. Of particular interest is the over-representation of a haplotype involving marker 2 allele 5 and the short allele of the LPR, due to the functional implications of the LPR alleles. Furthermore, marker 2 allele 5 is consistently present in haplotypes that give the strongest signals based on the largest number of observations.

Table 4.3.20. Summary of positive and near positive association results for four marker haplotypes in 'sliding window'.

	TDT					HHRR			
	markers / alleles	T	χ^2	p		**markers /** alleles	T	χ^2	p
Ant	**10, 11, 12, 13 /**					**1 2 3 4 /**			
	2 2 2 2	+	6.25	0.012		1 6 5 1*		4.68	0.031
	11, 12, 13, 14 /								
	2 2 2 1	+	4.57	0.033					
CVCR	**1, 2, 3, 4 /**					**1, 2, 3, 4 /**			
	1 7 3 1	-	5.00	0.025		1 3 4 1*	-	3.81	0.051
	1 5 3 1	+	3.60	0.058		1 5 3 1	+	4.23	0.040
						1 7 3 1	-	4.39	0.036
						1 7 4 1*	-	6.28	0.012
						Global		22.01	0.037
						2, 3, 4, LPR /			
						3 4 1 2*	-	3.90	0.048
						5 3 1 1	+	4.59	0.052
						6 4 1 1*	-	4.71	0.030
						7 3 1 2*	-	4.43	0.035
						10, 11, 12, 13 /			
						1 1 2 2*	-	4.78	0.029
						11, 12, 13, 14 /			
						1 2 2 2*	-	4.76	0.029
						Global			0.051
Both	**4, LPR, 6, 7 /**					**1, 2, 3, 4 /**			
	1 2 7 1*	-	4.00	0.045		1 3 4 1*	-		0.060
	6, 7, 8, 9 /					1 5 3 1	+		0.020
	9 1 2 2*	+	4.00	0.045		1 5 5 1*	-		0.030
	10, 11, 12, 13 /					1 7 4 1*	-		0.020
	2 2 2 2	+	6.12	0.013		Global			0.064
						2, 3, 4, LPR /			
						3 4 1 2	-	3.23	0.072
						5 3 1 1	+	4.02	0.045
						7 4 1 1*	-	5.39	0.020
						10, 11, 12, 13 /			
						1 1 2 2	-	4.64	0.031
						11, 12, 13, 14 /			
						1 2 2 2	-	4.73	0.030

* p values based on < 5 observations. T: Transmission (over, +, or under, -.); p: p value; TDT: transmission disequilibrium test; HHRR: haplotype-based haplotype relative risk.

Figure 4.3.15. Single marker and 'sliding window' associations in Antioquia.

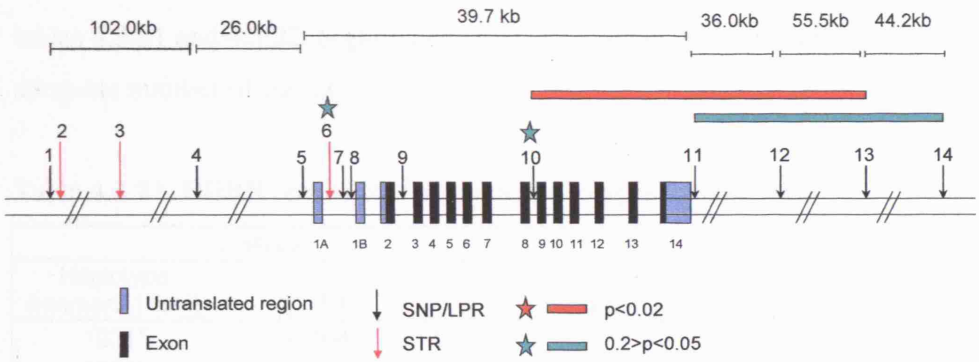
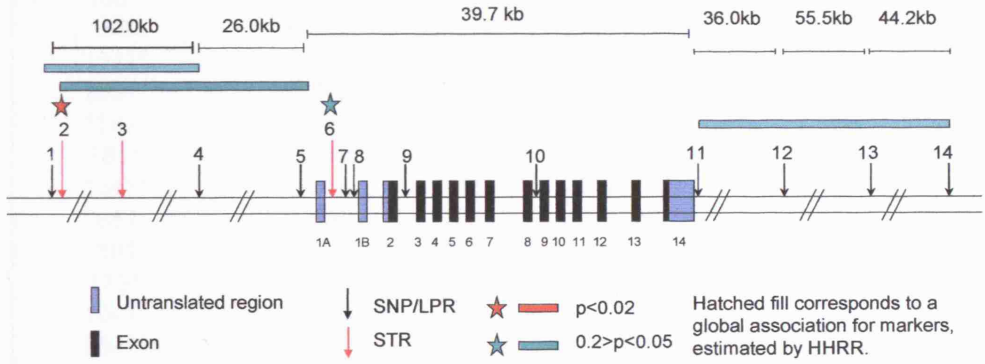
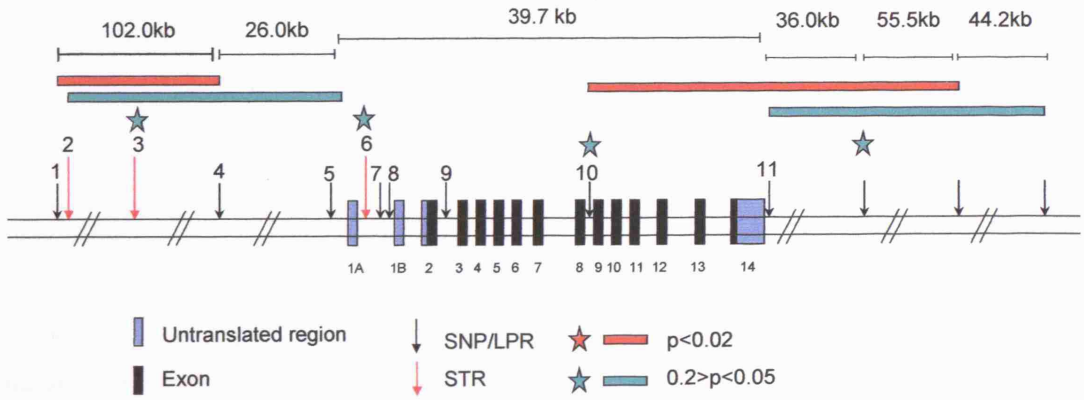


Figure 4.3.16. Single marker and 'sliding window' associations in CVCR.



p value for the global test of association for markers 11, 12, 13 and 14 is just outside conventional significance at 0.051.

Figure 4.3.17. Single marker and 'sliding window' associations in combined sample.



4.3.3.4 LD Block Haplotypes

Results of association analyses on the LD block haplotypes are presented in tables 4.3.21 and 4.3.22. Significance was not reached for any haplotype with an adequate number of transmissions.

Table 4.3.21. HHRR results for LD block haplotypes in Antioquia.

5' Block			3' Block		
Haplotype (markers 1 to 4)	χ^2 (1df)	p	Haplotype (markers 7 to 14)	χ^2 (1df)	p
1231*	0.064	0.800	12111212*	0.350	0.554
1331*	0.044	0.834	12211212	0.141	0.707
1341*	0.286	0.593	12211222*	0.092	0.761
1421*	0.063	0.801	12212221*	0.345	0.557
1431*	0.128	0.721	21111212*	0.835	0.361
1511*	1.624	0.203	21121112*	0.612	0.434
1521*	1.250	0.264	21122211*	1.318	0.251
1531	1.337	0.247	21122221	0.411	0.522
1532	0.080	0.777	21211212*	0.885	0.347
1541*	0.340	0.560	22111112	0.457	0.499
1551*	2.742	0.098	22111212*	0.003	0.954
1562*	1.057	0.304	22111222*	1.368	0.242
1631	0.248	0.618	22112221*	0.059	0.807
1641*	2.018	0.155	22121212*	1.857	0.173
1651*	4.680	0.031	22122221*	1.311	0.252
1661*	1.808	0.179			
1731	0.078	0.781			
1941*	0.702	0.402			
2541*	0.777	0.378			
2551	1.074	0.300			
2641*	0.129	0.719			
2661*	0.501	0.479			
Global test (haps >1%)	12.582 (16 df)	0.703	Global test (haps >1%)	7.834 (10 df)	0.645

* p values based on < 5 observations.

Table 4.3.22. HHRR results for LD block haplotypes in CVCR.

5' Block			3' Block		
Haplotype (markers 4 to 5)	χ^2 (1df)	p	Haplotype (markers 7 to 14)	χ^2 (1df)	p
11	0.104	0.747	11222121*	2.209	0.137
12	0.039	0.843	12111211*	2.856	0.091
21*	1.970	0.160	12211112*	0.614	0.433
22	0.009	0.924	12211212	0.135	0.714
			12211222*	1.533	0.216
			12212222*	0.791	0.374
			12222221*	1.477	0.224
			21122221	0.923	0.337
			21122222*	1.459	0.227
			22111112	0.037	0.847
			22111212	1.824	0.177
			22111222*	3.131	0.077
			22112221*	0.147	0.701
			22112222*	0.673	0.412
			22121212*	0.067	0.796
			22211212*	0.676	0.411
Global test on haps >1%	2.121 (3 df)	0.548		9.716 (8 df)	0.286

* p values based on < 5 observations.

4.4 Discussion

In this study I was concerned with comprehensively defining the variation present in the *SLC6A4* gene in BPI trios from two Latin American populations using five intragenic SNPs, five flanking SNPs, the functional promoter variant and three STRs, that span approximately 303.5kb around the gene. These markers were selected on the basis of minor allele frequencies, locus heterozygosities and physical locations, to enable the identification of *SLC6A4* regions that may be important in bipolar affective disorder.

4.4.1 Usefulness of Markers Selected

As relatively little information existed in public databases for some of these loci, markers were not equally informative. For example, the two 5' flanking SNPs had minor allele frequencies of less than 10% in both populations. The three STRs I used were also not of equal value; marker 3 was not as informative as marker 2 or marker 6, as it was the least polymorphic and deviated from HWE in Antioquia. Conversely markers 2 and 6, located 127,028bp apart in the upstream region of the SERT gene, amplified well and obeyed HWE. Marker 6 in particular showed a high level of heterozygosity and, due to its location, could be a marker of choice in future association studies. Although careful assessment of heterozygosities should be made before large-scale genotyping is undertaken, the usefulness of screening publicly available databases to find markers for gene mapping studies, made possible by the accessibility of the human genome sequence, has been demonstrated here.

4.4.2 Population Allele Frequencies and Heterozygosities

Characterising the study markers has thrown more light on the genetic relationship between Antioquia and CVCR. Some small fluctuation between populations in allele frequencies has been demonstrated in the SNP data. Notably, however, those SNPs with the rarest minor allele frequencies in Antioquia occur at similar frequencies in the CVCR; and when chi-square analyses is performed on the distribution of alleles between populations no p value falls below 1. Although

Antioquia shows a slightly higher average heterozygosity over the bi-allelic markers than CVCR, this is not statistically significant.

I found the frequencies of the LPR alleles to be 0.46 (S) / 0.54 (L) in Antioquia, and 0.54 (S) / 0.46 (L) in CVCR. Frequencies of LPR alleles for several global populations are shown in table 4.4.1. From this it can be seen that Antioquia is most similar to Sardinia and a large mixed caucasian sample (Lesch et al., 1996; Piccardi et al., 2002); whereas values for CVCR approach those found in the Maya population. This observation further confirms the strong European autosomal component to Antioquia, as indicated in Chapter 3.

Table 4.4.1. S and L allele frequencies in global populations.

Population	n	LPR allele S	LPR allele L	Author
Antioqua	**96**	**0.463**	**0.537**	Present study
CVCR	**107**	**0.545**	**0.455**	Present study
Caucasian (non-hispanic)	505	0.430	0.570	(Lesch et al., 1996)
Sardinia	202	0.460	0.540	(Piccardi et al., 2002)
Mbuti (South Africa)	36	0.111	0.889	(Gelernter et al., 1999)
Chinese	45	0.700	0.289	(Gelernter et al., 1999)
Naisoi (Australo-Melanesian)	19	0.289	0.711	(Gelernter et al., 1999)
Maya (Yucatan)	38	0.605	0.395	(Gelernter et al., 1999)
Rondonian Surui (Amazon Basin)	24	0.354	0.646	(Gelernter et al., 1999)

Heterozygosities of the STRs are also very similar in the two populations with CVCR being slightly more heterozygous over these markers. However, Antioquia has more marker 6 alleles than CVCR, and has a marginally higher mean heterozygosity when marker 3 is ignored (which deviated from the Hardy Weinberg equilibrium in Antioquia). Differences in STR heterozygosity do not reach statistical significance between the populations.

Overall, allele frequencies and heterozygosities are relatively similar between the two populations and most likely mirror their parallel demographic histories. These findings support previous research that showed Antioquia and CVCR to be genetically similar (Carvajal-Carmona et al., 2003). Considering LPR and SNP data together, CVCR may be demonstrating a slightly weaker European contribution at this gene. Although my conclusions are drawn from the analysis of one gene, and other genes may suggest alternative theories as a result of different gene histories (Jorde et al., 2000; Zhivotovsky et al., 2003), this evidence combined with that of Carvajal *et al*

(2003) confirms the strong genetic similarity of these populations. This similarity may be exploitable in fine-mapping and association studies (Carvajal-Carmona et al., 2003).

4.4.3 Linkage Disequilibrium

Part of this study was concerned with illuminating LD structure around *SLC6A4* in Antioquia and CVCR, so population comparisons could be made of LD patterns in non-transmitted chromosomes, often referred to as background LD (Freimer et al., 1997), and contrasted with patterns in chromosomes transmitted to BPI affecteds. Haplotypes were generated, non-transmitted and transmitted chromosomes separated from one another and statistical significance of allelic association estimated using FETp values, while the strength of LD was measured using Lewontin's D'.

4.4.3.1 Background Linkage Disequilibrium

Both measures defined a large 3' block of strong LD stretching from marker 7 in intron 1A to marker 14, located approximately 135kb downstream from the 3' UTR. Additionally, both approaches showed a less well defined block in the CVCR than Antioquia. Although the more distal markers would be expected to reveal the effects of recombination first, this pattern could represent moderate decay of this LD block in CVCR. There is the possibility that the 3' block of LD extends considerably further in the 3' direction. It is important to determine exactly how far as the coding region of *SLC6A4* could be in LD with a sizeable region of chromosome 17, and might therefore be important knowledge for LD mapping studies. Data from the HapMap project suggests that LD may stretch as far as 500kb downstream in the European CEPH, but perhaps less far in the African and Eastern Asian populations.

The structure of background LD 5' to *SLC6A4* is harder to define. According to FET p values, LD in the large block breaks down by marker 6 with the exception of some LD maintained with the LPR, and some spurious cases involving the 5' SNPs. D' values, on the other hand, clearly show the end of the LD block by marker 6 and do not suggest strong LD between the LPR and *SLC6A4* coding sequence. Reasons for this discrepancy may lie in the attributes of the two different types of test. For example, weak LD can still be statistically significant using FET, particularly in large

samples. Additionally, D' values that are neither 1 nor 0 are hard to interpret correctly (Ardlie et al., 2002). This implies that low D' values do not necessarily reflect the absence of LD. Together these results may indicate that the LPR is in LD with the SERT gene, but this LD is not perfect.

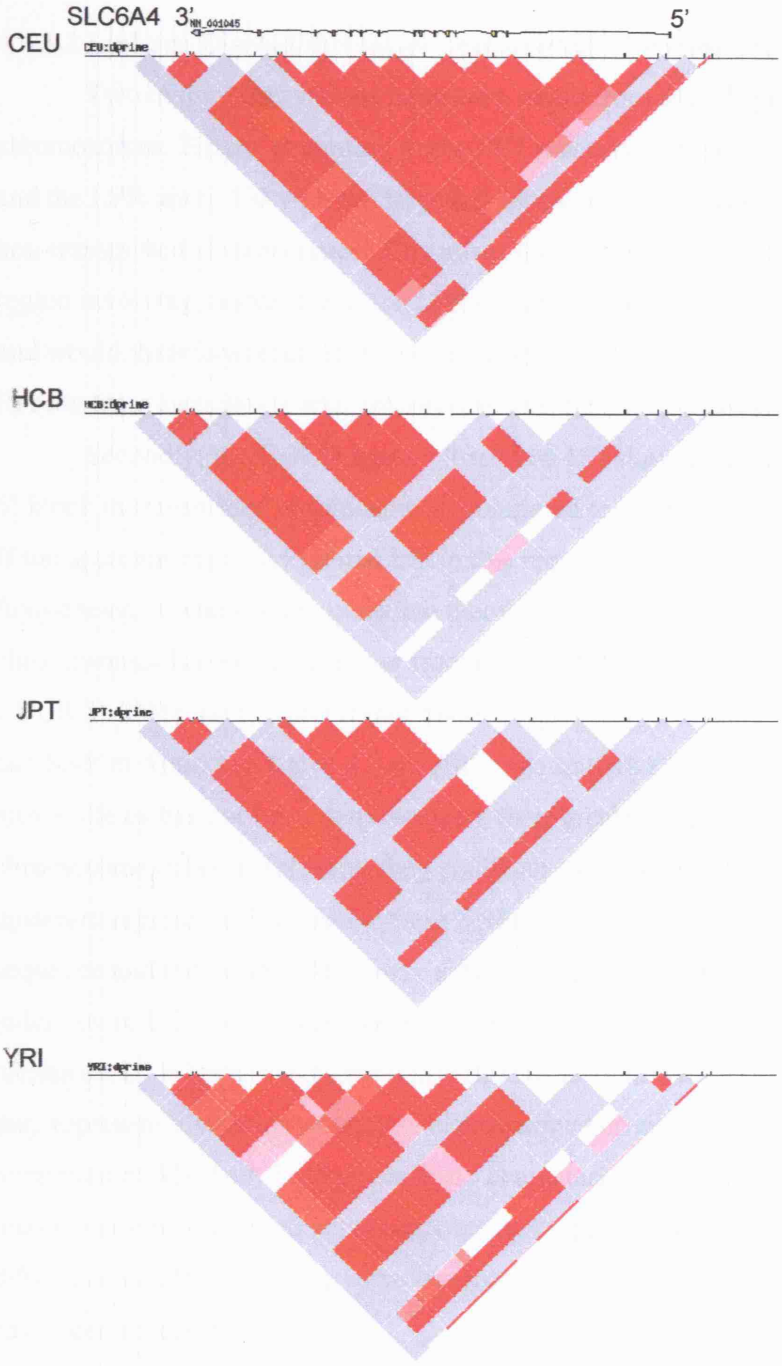
D' results show there to be an upstream LD block from around the LPR extending 5' to marker 1 in both populations. This two-block structure is most clearly illustrated in Antioquia (figure 4.3.6), although it is not well defined using FET. On closer inspection all of the high pairwise D' values in the 5' block involve one of the two 5' SNPs with very low minor allele frequencies. It is probable this block is reflective of this, as D' values are inflated by rare alleles (Ardlie et al., 2002).

Overall there is a less consistent pattern of LD in the 5' region of *SLC6A4* than the rest of the gene. This may be partly explained by the characteristics of the markers in the upstream region, and reflect different influences on estimates of LD dependent on the statistical test used. For example, two multi-allelic STRs and two rare minor allele SNPs were included which may inflate D' values (Ardlie et al., 2002; Teare et al., 2002).

When this region is compared to the HapMap Project the general LD pattern seen here is repeated in all four reference populations (European, Chinese, Japan, Nigeria); that is, relatively strong and extensive LD around most of *SLC6A4* and downstream sequence, some LD in the upstream regions and a break in LD in the vicinity of the LPR. Absolute levels differ between all populations, which is expected as they are likely to have diverse demographic histories, and demographic history is known to influence population LD patterns (Laan and Paabo, 1997; Reich et al., 2001). The CEPH population of European descent most closely mirrors the patterns of background LD in both of my populations, further validating my findings and the European contribution to these populations.

One cause for the lower 5' LD compared to the remainder of the gene, in both Antioquia and CVCR, might be a shared demographic history. On the other hand, the fact that LD blocks appear to border the LPR in both populations, an observation repeated in the demographically distinct HapMap populations, suggests there is something specific to the sequence surrounding the LPR that confers either high recombination or mutation rates; thereby diminishing LD. Therefore, one

Figure 4.4.1 LD figures around *SLC6A4* in the four HapMap Project reference populations.



Ticks represent markers and are positioned with respect to *SLC6A4* (top of figure). LD is measured by D' and strength of pairwise LD between markers is indicated by shade of red. Light grey represents D' not calculated or available. CEU: CEPH (Utah residents with ancestry from northern and western Europe); HCB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; YRI: Yoruba in Ibadan, Nigeria.

interpretation of these results is that a recombination hotspot in the vicinity of the LPR has been identified.

4.4.3.2 Linkage Disequilibrium in Transmitted Chromosomes

Two major observations have been made from the LD present in transmitted chromosomes. Firstly, according to the FET p results it appears that both marker 6 and the LPR are in LD with the large LD block, whereas marker 6 is not in LD in the non-transmitted chromosomes. This raises the question of whether this extended LD region involving marker 6 and the LPR confers a genetic predisposition to disease, and would therefore result in over-representation of marker 6 haplotypes in affecteds. Results from association analysis support this theory, discussed in more detail later.

Secondly, the GOLD figures (based on D') show a decrease in LD in the small 5' block in transmitted chromosomes, compared to their non-transmitted counterparts. If the apparent high background LD in this region is due to rare minor allele frequencies, it would seem plausible that the reduction in LD in the transmitted chromosomes is caused by higher frequencies of these minor alleles. Both transmitted CVCR 5' SNPs have higher minor allele frequencies, and this trend is repeated for one SNP in Antioquia (table 4.3.4). Although statistical over-transmission of the minor alleles has not been demonstrated, their greater frequency in the transmitted chromosomes may be relevant; they could act as markers for functionally important upstream regions or, less likely, these SNPs may lie in functionally important sequence and the minor alleles may act as disruptive mutations. Alternatively, if the reduction in LD in case chromosomes is due to increased recombination (of the LD measures, D' is known to be most sensitive to recombination (Teare et al., 2002)) it may represent disruption to healthy chromosomal arrangement, important in the regulation of *SLC6A4*. Further sequence characterisation and assessment of recombination is required here to answer these questions. Nonetheless, an important difference in LD patterns between transmitted and non-transmitted chromosomes may have been revealed.

4.4.3.3 LD Block Structure

Entering parental and index genotype data into HaploView confirmed a two block structure in all chromosome groups except the Antioquian parents, for whom no 5' block was identified. In general results provide good agreement with other LD

results; a two block structure has again been demonstrated and higher LD around the LPR has been shown in case chromosomes. One exception is the absence in HaploView of a 5' block in the Antioquian parental chromosomes (figure 4.3.10) and this discrepancy may be due to the fact that marker 2 is excluded in HaploView so the 5' block here is based only on the two 5' SNPs; whereas marker 2 is included in the 5' block using GOLD.

Interestingly, the LPR is part of the 5' block in CVCR for both parental and case chromosomes, whereas it is not part of any block in Antioquia. If the sequence around the LPR is part of a recombination hotspot the inclusion/exclusion of the LPR in an LD block may suggest there is no specific recognition site for chromosomal breakage, but rather recombination occurs across a specific genomic region. Although determinants underlying the physical locations of crossing-over events and recombination hotspots are still not well understood (Lupski, 2004), my observation agrees with evidence that shows crossing-over events to be normally distributed within hotspots that span regions as narrow as 1 to 2 kb (Jeffreys et al., 2001; Kauppi et al., 2004; Schneider et al., 2002).

The definition of block haplotypes has enabled the distribution of haplotypes to be assessed between chromosome groups (i.e Antioquia transmitted/nontransmitted and CVCR transmitted/nontransmitted). In all groups of chromosomes the same three 3' block haplotypes, generated from markers 10 to 14, predominated and further indicates the genetic similarity of the two populations. Little difference was seen between parental and case chromosomes, whereas the most notable difference in haplotype distribution occurred between all Antioquia chromosomes and all CVCR chromosomes. This suggests a lack of involvement of this part of *SLC6A4* in BPI. Haplotypes were limited to just five markers, however, and inclusion of more 3' markers may reveal an alternative scenario. Of more interest would be the comparison of 5' block haplotype distribution, which was not possible as this block is less well conserved between the groups. However, all blocks were tested separately for association to BPI, as discussed later.

In summary, the analysis of LD in *SLC6A4* has shown this gene to be in strong LD with a comparatively large region of chromosome 17. LD patterns in DNA 5' to the coding sequence are less well defined, and a hotspot of recombination may have been revealed in this region with its borders lying somewhere either side of the LPR.

A difference in LD pattern between transmitted and non-transmitted *SLC6A4* copies has also been illustrated, highlighting the upstream regions for future study. To determine what exactly is responsible for the observed difference in LD patterns in the 5' block – allele frequencies or recombination - requires further testing and may have important implications. For example, the sequence surrounding the 5' markers could be studied in more detail to find out if it includes important motifs such as transcription factor binding sites and enhancer recognition motifs, and if so whether a decay in LD may indicate an alteration to normal sequence and/or genomic arrangement. Transcription factor binding in healthy and unhealthy chromosomes could also be analysed using, for example, footprinting-based experiments.

This work illustrates the value of distinguishing between LD that occurs in normal populations and LD around disease alleles, and I would recommend using such methodology in future association studies.

4.4.4 Association

Due largely to its biological function, implication of the serotonin neurotransmitter in human behaviour and mood and its role in antidepressant drug activity (Hariri et al., 2002; Lucki, 1998; Ozaki et al., 2003), a lot of work has been done to associate *SLC6A4* with bipolar affective disorder. Some studies have proved successful (Battersby et al., 1996; Collier et al., 1996; Ogilvie et al., 1996; Rees et al., 1997), although associations are very rarely repeated. Many more have shown no statistical association (Bocchetta et al., 1999; Esterling et al., 1998; Geller and Cook, 1999; Ospina-Duque et al., 2000). Reasons are likely due to the genetic heterogeneity of BP meaning different populations could have different susceptibility alleles, further compounded by phenotype class definitions differing between studies as different classes of phenotype may result from different genetic components (Escamilla et al., 1997). Good experimental design is therefore a particularly important consideration in BP association studies and may strongly influence the chances of finding a true association. Considering these factors, my study includes several features to increase the power of detecting true signals, as well as increase the robustness of results to further testing. For example, I have employed fourteen markers, including the functional LPR, which comprehensively span over 300kb around *SLC6A4*. I have used two different kinds of family based association test - the TDT and the HHRR –

on both single markers and multi-marker haplotypes. Furthermore, I have provided the opportunity for results to be repeated within the same study by using two different study populations; Antioquia and CVCR.

In general, many significant associations have been demonstrated, the majority of which are likely meaningless, spurious p values; an artefact of a large number of computations. Considering three groups (Antioquia, CVCR and both combined) have been tested for fourteen individual loci and multi-marker haplotypes, in the TDT and HHRR, a certain number of significant associations may be expected by chance, due to multiple testing, amongst the vast majority of non-significant results.

Considering this, the significant associations that were consistently repeated may be the most interesting. In particular, marker 6 allele 9 is significantly over-represented in three instances. Marker 10 allele 2 is over-represented in two instances, while marker 2 allele 5 is also over-represented in two separate instances. The 'sliding window' four marker haplotype analysis shows a haplotype involving markers 10/11/12/13 has been over-represented twice. Two different 1/2/3/4 haplotypes have been implicated repeatedly, as well as two 2/3/4/LPR haplotypes. In fact, alleles 5/3/1 of markers 2/3/4 have been over-represented in four separate instances. This suggests that the 5/3/1 haplotype may represent a 5' upstream chromosomal arrangement important in BP, adding to the implication of the 5' region by the LD analysis.

By concentrating association attempts on those haplotypes that make up the LD blocks the chances of generating false positives by multiple testing are reduced. A similar approach was used successfully by other researchers who separated LD around *SLC6A4* into blocks and highlighted regions of the gene that may be important in ADHD (Curran et al., 2005). Here, none of the haplotypes representing either of the LD blocks showed significant over or under transmission to BPI cases and therefore support for involvement of *SLC6A4* regions in BPI has not been provided. Other factors may contribute to these negative results however. For example, HaploView was only able to compute bi-allelic loci meaning the informative multi-allelic STRs were not considered when LD blocks were estimated; therefore the potential involvement of haplotypes including marker 6 was not assessed. Results for this marker from other analyses have been interesting. Also, markers 2 and 3 were added to association analysis after the Antioquia 5' block (involving markers 1 and 4) was generated. If these markers were able to be included in the estimation of LD blocks, the block structure may have been different. Additionally, analysis was restricted to

one test only (the HHRR) due to the number of markers involved in the haplotypes and the data handling capacity of Genehunter.

In summary, no multi-marker haplotype showed particularly strong association with LD. Any involvement of the 3' block seems unlikely as the positive results from the sliding window were not repeated when only the LD block haplotypes were considered, and all markers downstream from marker 7 are in LD. Association of the 5' region is less obvious though, as relatively strong global associations were achieved in two instances in the sliding window analysis for this region, combined with the inability of the LD block association analysis to confirm or reject these findings.

4.4.4.1 The LPR

My results have shown there to be no association between the LPR promoter polymorphism's short (14 repeat units) and long (16 repeat units) alleles with BPI, suggesting this locus does not affect this disorder. This agrees with the many negative results at this locus (Esterling et al., 1998; Geller and Cook, 1999; Ospina-Duque et al., 2000; Piccardi et al., 2002; Rees et al., 1997), but not with those studies that have showed association, including a recent meta-analysis (Collier et al., 1996; Furlong et al., 1998; Lasky-Su et al., 2005; Mynett-Johnson et al., 2000).

An alternative explanation for the lack of association may come from more recent research regarding the allelic heterogeneity of the LPR. For example, five other alleles have been reported with 15, 18, 19, 20 and 22 copies of the repeat unit (Gelernter et al., 1997; Lesch and Mossner, 1998; Nakamura et al., 2000). In addition, the original 14 and 16 repeat unit alleles have been further divided into four and six sequence variants, respectively (Nakamura et al., 2000). A lack of sensitivity to detect all LPR variants has therefore been used to explain the failure of some association studies (Ospina-Duque et al., 2000; Piccardi et al., 2002). As my LPR assay is only sensitive enough to distinguish between the original L and S alleles, association signals for one of the more recently characterised alleles would be diminished due to dilution by the other non-associated alleles.

Analysing haplotypes generated by the LPR and the most proximal STR (marker 6) may have resolved complications related to the allelic heterogeneity of the LPR, as a potentially important trend was observed. That is, haplotype 2/9 was significantly over-represented in transmitted chromosomes in four out of a possible

six tests (table 4.3.19). One of these values (TDT for both populations with dhskip on) was the most significant of all p values obtained so far at 0.0047. This result mirrors the over-representation of marker 6 allele 9 individually, and a role of this microsatellite is further supported by LD results. This locus was shown to be included in the large LD block in transmitted (but not non-transmitted) chromosomes by FET results (figures 4.3.3 and 4.3.5). Furthermore, association between LPR and marker 6 alleles is not statistically significant in non-transmitted chromosomes, but is in transmitted chromosomes (Table 4.3.18). Of interest, and in contrast to findings for the LPR of many researchers (Collier et al., 1996; Furlong et al., 1998; Heils et al., 1996; Mynett-Johnson et al., 2000), but not all (Piccardi et al., 2002), the over-represented haplotype here contains an LPR long allele.

One conclusion from this is that I have identified a rare class of LPR long allele in LD with allele 9 of marker 6, which is moderately associated with BPI. However, further characterisation of the LPR allele in affecteds with the 2-9 LPR-marker 6 haplotype is required, to assess two possibilities. One, the size of the associated LPR alleles needs confirmed as an allele with more than 16 repeat units may have been isolated. This could be done relatively quickly by repeating the same PCR assay, but using fluorescently labelled primers and detecting alleles with laser technology in an ABI Sequencer. Two, sequencing of the LPR allele could be done as it may be a sequence variant of the L allele that has been associated. If a specific susceptibility allele is identified then further tests of association and functionality could be done. Such findings would go some way to explaining why so many discrepancies have arisen in association studies based on the LPR. Irrespective of this, marker 6 characterised here may prove to be very helpful in identifying one of the newly characterised LPR alleles in future association studies of psychiatric disorders.

CHAPTER 5: DISCUSSION

CHAPTER 5: DISCUSSION

If biological complexity of all organisms on this planet is considered, humans would lie in the upper tail of the distribution. Ever since the early discoveries of Mendel and Darwin in the late 19th century it was realised that the cause for this complexity may lie in our genetic make-up, and led to a substantial amount of scientific research in the 20th century being dedicated to deciphering our genetics and mapping our genome. This work has culminated in the Human Genome Project; one of the most significant scientific achievements of the 20th and early 21st centuries which has made public the vast majority of sequence for the 3.08×10^9 base pairs of the human genome. Today the genome only awaits its finishing touches including the harder to sequence regions of the genome, such as tightly compacted heterochromatic DNA around the chromosomal centromeres (Collins et al., 2004).

The result of these efforts is that we now know that genes encode proteins, we know how this system operates and we have accurate estimates as to how many genes we have. Various features of the genome have also been made apparent including; the presence of large stretches of single copy, non-coding sequence; substantial amounts of duplicated and tandemly repeated DNA; various sources of genetic variation within the human genome; and the separation of the genome by recombination into blocks of LD.

This knowledge has led to better understanding of many biological fields including molecular, cellular and developmental biology, and of course medicine. Medical genetics is concerned with the genetic basis of human disease and amongst the major advances in human medicine a greater understanding of the genome and genetic machinery has facilitated, is the more recent developments in gene therapy techniques. Additionally, variation in the human genome can illuminate the evolutionary history of modern humans, and reveal demographic patterns that occurred during global colonisation by our species.

One area that has remained a bugbear for geneticists is the elucidation of the genetic components of common complex traits, and in particular complex disorders. Finding genes involved in these disorders is made difficult by their multi-factorial nature. For example, in a simple Mendelian trait a particular phenotype often corresponds to one particular gene. In complex traits, however, their heterogeneous nature means no direct relationship exists between phenotype and one causal factor. A

few different genes may give rise to the same phenotype. Alternatively, a larger set of smaller effect genes may be involved, and has led to two contrasting views of the genetic architecture of complex disorders: the common disease/common variant model where a few, common alleles give rise to a trait; and the genetic heterogeneity model where many alleles, each of small effect and low frequency, combine to produce a phenotype (Reich and Lander, 2001). In either case gene effects are less than in Mendelian traits. Furthermore, whatever the genetic background, the expression of a complex trait is likely to depend on an intricate relationship with the environment; the effect and size of which are hard to determine and remain largely unknown.

It is the multifactorial nature of complex disorders that allows them to attain high population prevalence. Since some individuals in a population will harbour some, but not all, susceptibility alleles required for a disease phenotype, their potentially deleterious effect will not be expressed, and these alleles will evade removal by selection. This will allow complex disorder alleles to reach relatively large population frequencies, and increases the probability of individuals carrying the combination of susceptibility alleles required for a disorder. Hence, many complex disorders are relatively common when compared to Mendelian diseases which are often rare.

Identifying susceptibility loci for common complex disorders is not straightforward and there exists a disproportionate lack of repeatable, positive results. A lack of understanding of the mechanisms underlying complex disorders and traits in general may contribute to failure as this could lead to poor experimental design. For example, factors such as the mode of inheritance of the risk allele, its penetrance and expected population allele frequencies are important considerations when trying to maximise the power of your experiment to detect a positive signal (Zondervan and Cardon, 2004). Furthermore, due to a heterogeneous background it is possible that different factors cause the same disorder in different individuals. This seriously compounds all strategies for finding susceptibility loci. Despite these complications some successes have been achieved for disorders with strong biological markers, such as Alzheimer's disease (Scacchi et al., 1999), type 2 diabetes (Altshuler et al., 2000) and bladder cancer (Engel et al., 2002).

One group of complex disorders with no such obvious biological markers are the psychiatric disorders, which adds another level of complication to gene discovery

attempts. These disorders manifest themselves in aberrant behaviour, personality and mood, and in their most extreme cases can be very debilitating. The environment is believed to be particularly influential and in some cases phenotypes may appear to be over-reactions to environmental stimuli. Without predefined physical markers, diagnosis relies heavily on the subjective opinions of doctors and psychiatrists, although recognised criteria (such as DSMIV) are adhered to as closely as possible and several independent diagnoses need to agree before treatment can be administered. Nonetheless, the possibility that the same diagnosis is given to slightly different phenotypes remains, and further reduces the chances of successful gene discovery.

In my research I have attempted to increase the potential of association studies to find genes for disorders such as BPAD by exploring in detail the distribution and functionality of human genetic variation. I have approached this in two alternative yet complementary ways: i. I have investigated the molecular variation in the promoters of genes involved in normal behaviour and mood, and assessed the functionality of this variation, thereby identifying a novel set of functional candidate loci; ii. I have studied the distribution of genetic diversity within the Antioquian population isolate thereby defining the genetics of this 'special' population, and simultaneously illuminating its evolutionary history. Finally, I have conducted a detailed association study in two Latin American population isolates, including a description of the distribution of genetic diversity around a neurotransmitter transporter in BPI affecteds and unaffecteds, to exclude this gene as a potential candidate gene for BPAD in these populations.

To date, studies interested in determining the genetics underlying human disorders have focused on coding mutations important to amino acid sequence or structure of polypeptides; however an alternative source of functional variation exists. Due in part to the observation that a large amount of coding genetic material is shared between humans and great apes, and that the differences may not be sufficient to explain the obvious phenotypic differences, the importance to human diversity of variation in genetic expression was suggested around 30 years ago (King and Wilson, 1975). This idea is reinforced by more recent, post-genome, discoveries that a lower number of genes exist in the human genome than early estimates predicted (Collins et al., 2004). Subsequently, there has been an increase in interest in the prevalence of

variable gene expression between individuals; substantial levels of naturally occurring, functionally relevant, heritable genetic variation within regulatory DNA have been confirmed (Cheung et al., 2003; Oleksiak et al., 2002; Yan et al., 2002). Phenotypic effects of these regulatory variants are likely to be less than for coding mutations and may exist at high frequencies in populations. As such, they make excellent functional candidates for complex traits and disorders. Therefore, a useful resource for genetic studies of psychiatric disorders would be a set of regulatory variants from relevant genes, which represent good functional candidates.

My results may have provided a robust set of disease candidate alleles from the promoters of two genes important in the serotonin neurotransmission pathway: *SLC6A4* (human serotonin transporter) and *SLC18A2* (vesicular monoamime transporter). The promoters of both of these genes were very polymorphic, and a large proportion of the variation tested in *in vitro* assays was shown to be functional. An explanation for this high level of diversity, both in terms of sequence and function, may be a low level of functional constraint on this sequence. For example, the core promoters of these genes may not be the major regulatory factors, but rather fine-tune genetic expression which is under the control of a major controlling locus located elsewhere. Recent research has suggested that eukaryotic gene expression may be controlled by *trans* acting loci and major universal control loci that are the regulators of several genes simultaneously (Morley et al., 2004). Nonetheless, this work has revealed functional promoter polymorphisms in genes important to the synthesis and regulation of the ubiquitous serotonin neurotransmitter; a bioamine involved in many functions central to healthy human development and behaviour, and these variants should serve as important candidate loci for future direct association studies of psychiatric disorders. What's more, results could increase the understanding of mechanisms underlying complex psychiatric disorders and psychiatric traits.

Interestingly two genes, *HTR1A* and *TPH2*, also important in the serotonin pathway, showed very low levels of promoter diversity and may indicate functional constraints. For example *HTR1A* has more than one role in regulating synaptic serotonin and deregulation of expression would affect serotonin levels in more than one way. *TPH2*, on the other hand, may be the rate-limiting enzyme in the synthesis of brain 5HT. It is likely therefore that functional mutations that arise in the promoters of these two genes have large phenotypic effects; if these genes were implied in a disorder screening the promoter sequence for mutations would be very worthwhile.

Identifying novel functional candidate variants, as done here, may provide geneticists with extra tools to find genes important in psychiatric disorders. However, how such genetic variation is distributed throughout a population also influences the effectiveness of an association study. The distribution of genetic variation within a population is shaped by the demographic and evolutionary history of that population, giving rise to population-specific genetic architecture and distributions of traits and phenotypes. Population isolates may be ideal for association studies due to the distribution of genetic variation within them. For example they can have low genetic diversity, be genetically homogeneous, have a high number of alleles and genomic regions identical by descent and they can show extensive LD. Isolates may therefore reduce the compounding effects of disease heterogeneity and their high frequency of shared ancestral DNA make them well suited to direct and indirect association analyses and LD mapping (Freimer et al., 1997; Hall et al., 2002; Sheffield et al., 1998; Varilo and Peltonen, 2004; Wright et al., 1999). However, it has also been acknowledged that different 'special' populations will possess these traits to variable extents and require separate genetic characterisation for most efficient use in association studies (Varilo and Peltonen, 2004).

Antioquia is a Latin American geographic isolate from North West Colombia and could be a very useful population for mapping disease genes. This study is the first to characterise in detail its genetic make-up at the autosomal level, which is important as it is in the autosome that the majority of disease genes will reside. Autosomal loci also have several advantages over uniparental inherited systems for detecting traces of historical demographic events. Several demographic events in its past suggest Antioquia may show low levels of genetic diversity and heterogeneity, for e.g. parental populations went through contractions at foundation and it is still a relatively young population of around twenty generations. Levels of LD may be large due to its young age and the nature of the admixture at founding, as it was founded by genetically distant populations. In contrast, my results do not show either low diversity or strong LD in Antioquia when compared to the large outbred Spanish population. When compared to the Ticuna (a Native Colombian population isolate) which has little historical genetic mixing or demographic expansion, Antioquia was consistently shown to be more diverse and have less LD. Interestingly, Antioquia and

Spain were not significantly differentiated from each other and showed a high level of autosomal haplotype sharing.

The high diversity seen here in Antioquia is likely due in part to the genetic distances between its parental populations at admixture, and perhaps also to the population expansion Antioquia has experienced in the last 100 to 120 years. Additionally, the Spanish sample used here originated from southern Spain whereas colonists in the 16th and 17th centuries likely came from various Spanish regions adding to the diversity within the Spanish parental population. The usefulness of Antioquia in association studies might therefore be dependent on the size of founding populations. Although diversity levels are high, the majority of the autosome derives from a Spanish sample, small in size compared to the total Spanish population. This should increase the numbers of genomic regions identical by descent in present day Antioquia (enhancing the power of indirect association analyses) and decrease disease heterogeneity (improving the chances of finding a common susceptibility allele in direct tests). As genomic LD does not extend particularly far I would not recommend using Antioquia in broad, genome-wide association analysis reliant on long range LD between a marker and disease allele. Instead, this population may be better suited to fine mapping studies, where distances between disease and marker alleles are much shorter. One possibility that hasn't been fully explored yet is the use of Antioquia in admixture mapping. However, due to the large predominance of Spain in the autosomal admixture, this method may not be particularly effective (Patterson et al., 2004). More work is needed for confirmation of the relative contribution of each founding population, and to determine the frequencies of marker alleles in the parental populations.

My research shows Antioquia to have a unique distribution of genetic diversity at the autosomal level, reflective of its demographic history, and has confirmed that Antioquia may be a useful population isolate for finding complex disorder mutations. Knowledge of the genetics of a population is important when considering how it can best be used in association studies. Here, I have made several considerations in my study to help determine whether *SLC6A4* has a role in bipolar affective disorder in Antioquia; i. I have used family-based association analyses, which have the benefit of avoiding population stratification and maintaining the power to detect genes of modest effect, ii. Fourteen markers, spanning no more than

150kb in either direction from the gene, were used to describe variation across this gene; iii. Only a narrow and severe class of BP was included, BPI, which further reduces compounding effects of disease heterogeneity. Furthermore, I have addressed historical problems of repeating positive results in complex disorder mapping by including another genetically similar Latin American population isolate, from the Central Valley of Costa Rica. Although very strong signals were not detected, this work revealed potentially interesting differences in sequence 5' to *SLC6A4* coding DNA between chromosomes transmitted and non-transmitted to BPI affecteds. This is seen in disparate LD patterns as well as over-representation of multi-marker haplotypes; however these results may be an artefact of allele frequencies and require confirmation. Interestingly, a well known promoter based functional polymorphism (5HTTLPR) was also included in analysis, and it was with a two marker haplotype involving the LPR and a nearby STR that the most significant association was achieved. One possible explanation for this result is that the proximal STR has identified a role for one of the many recently discovered LPR sub-alleles or allelic sequence variants in BPI.

Due to the large number of computations in association analyses such as these a number of significant results may be expected by chance alone. Therefore, it is important to include several means of detecting differences in genetic variation between affected and non-affected individuals within studies, under the assumption that true signals will be consistently detected across tests. Considering this, the significance of association between the LPR haplotype and BP is further emphasized. Although characterisation of LPR sequence and functionality is required in those individuals showing a positive association, the identification of a long allele variant suggests a relationship between upregulation of *SLC6A4* and bipolar affective disorder. This should lead to a decrease in synaptic serotonin and, if this result is confirmed, may further the understanding of the mechanisms underlying BP. Interestingly, this study has highlighted upstream regions and promoter based polymorphism, further implying the importance of mutations that influence gene expression in psychiatric disorders.

Over the last 100 years great advances have been made in understanding the genetics underlying human complexity. Although genes have been found for many simple, Mendelian traits, less is known about traits with a more complex genetic and

environmental architecture. Due to their high prevalence in populations there is a great need to fully understand the genetics of such traits, particularly one class of debilitating complex trait: common disorders. In this study I have investigated several areas of human genetic diversity and performed a comprehensive gene association analysis which may contribute to one of the biggest challenges that remains for human geneticists in the 21st century; elucidation of the underlying mechanisms of common complex disorders.

Bibliography

Abdolmaleky, H.M., Cheng, K.H., Russo, A., Smith, C.L., Faraone, S.V., Wilcox, M., Shafa, R., Glatt, S.J., Nguyen, G., Ponte, J.F., Thiagalingam, S. and Tsuang, M.T. (2005) Hypermethylation of the reelin (RELN) promoter in the brain of schizophrenic patients: A preliminary report. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, **134B**, 60-66.

Abecasis, G.R. and Cookson, W.O.C. (2000) GOLD - Graphical Overview of Linkage Disequilibrium. *Bioinformatics*, **16**, 182-183.

Aita, V.M., Liu, J.J., Knowles, J.A., Terwilliger, J.D., Baltazar, R., Grunn, A., Loth, J.E., Kanyas, K., Lerer, B., Endicott, J., Wang, Z.Y., Penchaszadeh, G., Gilliam, T.C. and Baron, M. (1999) A comprehensive linkage analysis of chromosome 21q22 supports prior evidence for a putative bipolar affective disorder locus. *American Journal of Human Genetics*, **64**, 210-217.

Akey, J.M., Zhang, G., Zhang, K., Jin, L. and Shriver, M.D. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805-1814.

Ali, S. and Hasnain, S.E. (2002) Molecular dissection of the human Y-chromosome. *Gene*, **283**, 1-10.

Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T.J., Daly, M., Groop, L. and Lander, E.S. (2000) The common PPAR gamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*, **26**, 76-80.

Alvarez, V.M. (1996) Poblamiento y poblacion en el valle de Aburra y Medellin, 1541–1951. In J.O., M. (ed.), *Historia de Medellin.* Suramericana., Medellin., p. pp 23–27.

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*. Washington DC: American Psychiatric Association. 4th ed.

Anderson, G.M. (2004) Peripheral and central neurochemical effects of the selective serotonin reuptake inhibitors (SSRIs) in humans and nonhuman primates: assessing bioeffect and mechanisms of action. *International Journal of Developmental Neuroscience*, **22**, 397-404.

Antequera, F. and Bird, A. (1993) Number of Cpg Islands and Genes in Human and Mouse. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 11995-11999.

Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **3**, 299-309.

ArisBrosou, S. and Excoffier, L. (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution*, **13**, 494-504.

Arndt, P.F., Hwa, T. and Petrov, D.A. (2005) Substantial regional variation in substitution rates in the human genome: Importance of GC content, gene density, and telomere-specific effects. *Journal of Molecular Evolution*, **60**, 748-U728.

Avise, J.C. (2004) *Molecular markers, natural history, and evolution.* Sinauer Associates Incorporated, Sunderland, Massachusetts.

Badenhop, R.F., Moses, M.J., Scimone, A., Mitchell, P.B., Ewen, K.R., Rosso, A., Donald, J.A., Adams, L.J. and Schofield, P.R. (2001) A genome screen of a

large bipolar affective disorder pedigree supports evidence for a susceptibility locus on chromosome 13q. *Molecular Psychiatry*, **6**, 396-403.

Bamshad, M.J., Mummidi, S., Gonzalez, E., Ahuja, S.S., Dunn, D.M., Watkins, W.S., Wooding, S., Stone, A.C., Jorde, L.B., Weiss, R.B. and Ahuja, S.K. (2002) A strong signature of balancing selection in the 5 ' cis- regulatory region of CCR5. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 10539-10544.

Barbujani, G., Magagni, A., Minch, E. and Cavalli-Sforza, L.L. (1997) An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 4516-4519.

Barnes, N.M. and Sharp, T. (1999) A review of central 5-HT receptors and their function. *Neuropharmacology*, **38**, 1083-1152.

Baron, M., Risch, N., Hamburger, R., Mandel, B., Kushner, S., Newman, M., Drumer, D. and Belmaker, R.H. (1987) Genetic-Linkage Between X-Chromosome Markers and Bipolar Affective-Illness. *Nature*, **326**, 289-292.

Baron, M., Straub, R.E., Lehner, T., Endicott, J., Ott, J., Gilliam, T.C. and Lerer, B. (1994) Bipolar Disorder and Linkage to Xq28. *Nature Genetics*, **7**, 461-461.

Barrantes, R., Smouse, P.E., Mohrenweiser, H.W., Gershowitz, H., Azofeifa, J., Arias, T.D. and Neel, J.V. (1990) Microevolution in Lower Central America - Genetic- Characterization of the Chibcha-Speaking Groups of Costa-Rica and Panama, and a Consensus Taxonomy Based on Genetic and Linguistic Affinity. *American Journal of Human Genetics*, **46**, 63-84.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263-265.

Barton, N.H. and Keightley, P.D. (2002) Understanding quantitative genetic variation. *Nature Reviews Genetics*, **3**, 11-21.

Battersby, S., Ogilvie, A.D., Smith, C.A.D., Blackwood, D.H.R., Muir, W.J., Quinn, J.P., Fink, G., Goodwin, G.M. and Harmar, A.J. (1996) Structure of a variable number tandem repeat of the serotonin transporter gene and association with affective disorder. *Psychiatric Genetics*, **6**, 177-181.

Bearden, C.E., Reus, V.I. and Freimer, N.B. (2004) Why genetic investigation of psychiatric disorders is so difficult. *Current Opinion in Genetics & Development*, **14**, 280-286.

Belkhir, K. GENETIX, logicel sous Windows TM pour la genetique des populations. Laboratoire Genome, Populations, Interactions CNRS UMR 5000, Universite de Montpellier II, Montpellier, France.

Bell, D.A. and Taylor, J.A. (1997) Genetic analysis of complex diseases. *Science*, **275**, 1327-1328.

Bellivier, F., Henry, C., Szoke, A., Schurhoff, F., Nosten-Bertrand, M., Feingold, J., Launay, J.M., Leboyer, M. and Laplanche, J.L. (1998) Serotonin transporter gene polymorphisms in patients with unipolar or bipolar depression. *Neuroscience Letters*, **255**, 143-146.

Berrettini, W.H., Ferraro, T.N., Goldin, L.R., Weeks, D.E., Deterawadleigh, S., Nurnberger, J.I. and Gershon, E.S. (1994) Chromosome-18 Dna Markers and Manic-Depressive Illness - Evidence For a Susceptibility Gene. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 5918-5921.

Bertina, R.M., Koeleman, B.P.C., Koster, T., Rosendaal, F.R., Dirven, R.J., Deronde, H., Vandervelden, P.A. and Reitsma, P.H. (1994) Mutation in Blood-

Coagulation Factor-V Associated with Resistance to Activated Protein-C. *Nature*, **369**, 64-67.

Blackwood, D.H.R., He, L., Morris, S.W., McLean, A., Whitton, C., Thomson, M., Walker, M.T., Woodburn, K., Sharp, C.M., Wright, A.F., Shibasaki, Y., StClair, D.M., Porteous, D.J. and Muir, W.J. (1996) A locus for bipolar affective disorder on chromosome 4p. *Nature Genetics*, **12**, 427-430.

Blangero, J. (2004) Localization and identification of human quantitative trait loci: King Harvest has surely come. *Current Opinion in Genetics & Development*, **14**, 233-240.

Bocchetta, A., Piccardi, M.P. and Delzompo, M. (1994) Is Bipolar Disorder Linked to Xq28. *Nature Genetics*, **6**, 224-224.

Bocchetta, A., Piccardi, M.P., Palmas, M.A., Oi, A. and Del Zompo, M. (1999) Family-based association study between bipolar disorder and DRD2, DRD4, DAT, and SERT in Sardinia. *American Journal of Medical Genetics*, **88**, 522-526.

Bonatto, S.L. and Salzano, F.M. (1997) Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the new world. *American Journal of Human Genetics*, **61**, 1413-1423.

Bortolini, M.C., Da Silva, W.A., De Guerra, D.C., Remonatto, G., Mirandola, R., Hutz, M.H., Weimer, T.A., Silva, M., Zago, M.A. and Salzano, F.M. (1999) African-derived south American populations: A history of symmetrical and asymmetrical matings according to sex revealed by bi- and uni-parental genetic markers. *American Journal of Human Biology*, **11**, 551-563.

Bortolini, M.C., Salzano, F.M., Thomas, M.G., Stuart, S., Nasanen, S.P.K., Bau, C.H.D., Hutz, M.H., Layrisse, Z., Petzl-Erler, M.L., Tsuneto, L.T., Hill, K., Hurtado, A.M., Castro-de-Guerra, D., Torres, M.M., Groot, H., Michalski, R., Nymadawa, P., Bedoya, G., Bradman, N., Labuda, D. and Ruiz-Linares, A. (2003) Y-chromosome evidence for differing ancient demographic histories in the Americas. *American Journal of Human Genetics*, **73**, 524-539.

Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, **33**, 228-237.

Boularand, S., Darmon, M.C., Ravassard, P. and Mallet, J. (1995) Characterization of the Human Tryptophan-Hydroxylase Gene Promoter - Transcriptional Regulation by Camp Requires a New Motif Distinct from the Camp-Responsive Element. *Journal of Biological Chemistry*, **270**, 3757-3764.

Bowcock, A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K. and Cavalli-Sforza, L.L. (1991) Drift, Admixture, and Selection in Human-Evolution - a Study with DNA Polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 839-843.

Bowcock, A.M., Ruizlinares, A., Tomfohrde, J., Minch, E., Kidd, J.R. and Cavalli-Sforza, L.L. (1994) High-Resolution of Human Evolutionary Trees with Polymorphic Microsatellites. *Nature*, **368**, 455-457.

Bray, N.J., Buckland, P.R., Williams, N.M., Williams, H.J., Norton, N., Owen, M.J. and O'Donovan, M.C. (2003) A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain. *American Journal of Human Genetics*, **73**, 152-161.

Breidenthal, S.E., White, D.J. and Glatt, C.E. (2004) Identification of genetic variants in the neuronal form of tryptophan hydroxylase (TPH2). *Psychiatric Genetics*, **14**, 69-72.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J. and Rolf, B. (1998) Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *American Journal of Human Genetics*, **62**, 1408-1415.

Brown, T.A. (2002) *Genomes*. Bios Scientific Publishers Ltd, Oxford, U.K.

Brunet, M., Guy, F., Pilbeam, D., Mackaye, H.T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J.R., De Bonis, L., Coppens, Y., Dejax, J., Denys, C., Duringer, P., Eisenmann, V.R., Fanone, G., Fronty, P., Geraads, D., Lehmann, T., Lihoreau, F., Louchart, A., Mahamat, A., Merceron, G., Mouchelin, G., Otero, O., Campomanes, P.P., De Leon, M.P., Rage, J.C., Sapanet, M., Schuster, M., Sudre, J., Tassy, P., Valentin, X., Vignaud, P., Viriot, L., Zazzo, A. and Zollikofer, C. (2002) A new hominid from the Upper Miocene of Chad, central Africa. *Nature*, **418**, 145-151.

Buckland, P.R., Hoogendoom, B., Guy, C.A., Coleman, S.L., Smith, S.K., Buxbaum, J.D., Haroutunian, V. and O'Donovan, M.C. (2004) A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. *Biochimica Et Biophysica Acta-Molecular Basis of Disease*, **1690**, 238-249.

Burenhult, G. (1993) *The First humans : human origins and history to 10,000 BC*. HarperSanFrancisco, San Francisco.

Cadoret, R.J., Yates, W.R., Troughton, E., Woodworth, G. and Stewart, M.A. (1995) Genetic-Environmental Interaction in the Genesis of Aggressivity and Conduct Disorders. *Archives of General Psychiatry*, **52**, 916-924.

Cann, R.L., Stoneking, M. and Wilson, A.C. (1987) Mitochondrial-DNA and Human-Evolution. *Nature*, **325**, 31-36.

Cardno, A.G., Marshall, E.J., Coid, B., Macdonald, A.M., Ribchester, T.R., Davies, N.J., Venturi, P., Jones, L.A., Lewis, S.W., Sham, P.C., Gottesman, II, Farmer, A.E., McGuffin, P., Reveley, A.M. and Murray, R.M. (1999) Heritability estimates for psychotic disorders - The Maudsley Twin psychosis series. *Archives of General Psychiatry*, **56**, 162-168.

Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nature Reviews Genetics*, **2**, 91-99.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q. and Lander, E.S. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, **22**, 231-238.

Carvajal-Carmona, L.G., Ophoff, R., Service, S., Hartiala, J., Molina, J., Leon, P., Ospina, J., Bedoya, G., Freimer, N. and Ruiz-Linares, A. (2003) Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Human Genetics*, **112**, 534-541.

Carvajal-Carmona, L.G., Soto, I.D., Pineda, N., Ortiz-Barrientos, D., Duque, C., Ospina-Duque, J., McCarthy, M., Montoya, P., Alvarez, V.M., Bedoya, G. and Ruiz-Linares, A. (2000) Strong Amerind/white sex bias and a possible sephardic contribution among the founders of a population in northwest Colombia. *American Journal of Human Genetics*, **67**, 1287-1295.

Cavalli-Sforza, L.L. (1966) Population Structure and Human Evolution. *Proceedings of the Royal Society Series B-Biological Sciences*, **164**, 362-&.

Cavalli-Sforza, L.L. (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton, New Jersey.

Cavalli-Sforza, L.L. (2005) The Human Genome Diversity Project: past, present and future. *Nature Reviews Genetics*, **6**, 333-340.

Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J. and Deka, R. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 1041-1046.

Chakraborty, R. and Weiss, K.M. (1988) Admixture as a Tool for Finding Linked Genes and Detecting That Difference from Allelic Association between Loci. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 9119-9123.

Chen, W.M. and Deng, H.W. (2001) A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genetic Epidemiology*, **21**, 53-67.

Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics*, **33**, 422-425.

Clayton, D. (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *American Journal of Human Genetics*, **65**, 1170-1177.

Clayton, D. and Jones, H. (1999) Transmission/disequilibrium tests for extended marker haplotypes. *American Journal of Human Genetics*, **65**, 1161-1169.

Collier, D.A., Arranz, M.J., Sham, P., Battersby, S., Vallada, H., Gill, P., Aitchison, K.J., Sodhi, M., Li, T., Roberts, G.W., Smith, B., Morton, J., Murray, R.M., Smith, D. and Kirov, G. (1996) The serotonin transporter is a potential susceptibility factor for bipolar affective disorder. *Neuroreport*, **7**, 1675-1679.

Collier, D.A. and Sham, P.C. (1998) The usual suspects: tyrosine hydroxylase and the serotonin transporter in affective disorders. *Molecular Psychiatry*, **3**, 103-105.

Collins, F.S., Brooks, L.D. and Chakravarti, A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, **8**, 1229-1231.

Collins, F.S., Lander, E.S., Rogers, J. and Waterston, R.H. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.

Collins, F.S., Morgan, M. and Patrinos, A. (2003) The human genome project: Lessons from large-scale biology. *Science*, **300**, 286-290.

Collins-Schramm, H.E., Chima, B., Operario, D.J., Criswell, L.A. and Seldin, M.F. (2003) Markers informative for ancestry demonstrate consistent megabase-length linkage disequilibrium in the African American population. *Human Genetics*, **113**, 211-219.

Collins-Schramm, H.E., Phillips, C.M., Operario, D.J., Lee, J.S., Weber, J.L., Hanson, R.L., Knowler, W.C., Cooper, R., Li, H.Z. and Seldin, M.F. (2002) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *American Journal of Human Genetics*, **70**, 737-750.

Comings, D.E. and Okada, T.A. (1976) Fine-Structure of Heterochromatin of Kangaroo Rat Dipidomys- Ordii, and Examination of Possible Role of Actin and Myosin in Heterochromatin Condensation. *Journal of Cell Science*, **21**, 465-477.

Contente, A., Dittmer, A., Koch, M.C., Roth, J. and Dobbelstein, M. (2002) A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nature Genetics*, **30**, 315-320.

Cooke, G.S. and Hill, A.V.S. (2001) Genetics of susceptibility to human infectious disease. *Nature Reviews Genetics*, **2**, 967-977.

Cooke, R. (1998) Human settlement of Central America and northern most South America (14,000-8000 BP). *Quaternary International*, **50**, 177-190.

Cooper, G., Rubinsztein, D.C. and Amos, W. (1998) Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. *Human Molecular Genetics*, **7**, 1425-1429.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L. and Pericakvance, M.A. (1993) Gene Dose of Apolipoprotein-E Type-4 Allele and the Risk of Alzheimers-Disease in Late-Onset Families. *Science*, **261**, 921-923.

Cowles, C.R., Hirschhorn, J.N., Altshuler, D. and Lander, E.S. (2002) Detection of regulatory variation in mouse genes. *Nature Genetics*, **32**, 432-437.

Csink, A.K. and Henikoff, S. (1998) Something from nothing: The evolution and utility of satellite repeats. *Trends in Genetics*, **14**, 200-204.

Curran, S., Purcell, S., Craig, I., Asherson, P. and Sham, P. (2005) The serotonin transporter gene as a QTL for ADHD. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, **134B**, 42-47.

Curtis, D., Brynjolfsson, J., Petursson, H., Holmes, S., Sherrington, R., Brett, P., Rifkin, L., Murphy, P., Moloney, E., Melmer, G. and Gurling, H.M.D. (1993) Segregation and Linkage Analysis in 5 Manic-Depression Pedigrees Excludes the 5ht1a Receptor Gene (Htr1a). *Annals of Human Genetics*, **57**, 27-39.

Daly, M.J., Rioux, J.D., Schaffner, S.E., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nature Genetics*, **29**, 229-232.

Darvasi, A. and Shifman, S. (2005) The beauty of admixture. *Nature Genetics*, **37**, 118-119.

Das Gupta, R. and Guest, J.F. (2002) Annual cost of bipolar disorder to UK society. *British Journal of Psychiatry*, **180**, 227-233.

Davies, J.L., Kawaguchi, Y., Bennett, S.T., Copeman, J.B., Cordell, H.J., Pritchard, L.E., Reed, P.W., Gough, S.C.L., Jenkins, S.C., Palmer, S.M., Balfour, K.M., Rowe, B.R., Farrall, M., Barnett, A.H., Bain, S.C. and Todd, J.A. (1994) A Genome-Wide Search for Human Type-1 Diabetes Susceptibility Genes. *Nature*, **371**, 130-136.

de Jaramillo, G. and Vega, L. (2001) *Breve Historia De Antioquia*. Universidad de Antioquia.

Degn, B., Lundorf, M.D., Wang, A., Vang, M., Mors, O., Kruse, T.A. and Ewald, H. (2001) Further evidence for a bipolar risk gene on chromosome 12q24 suggested by investigation of haplotype sharing and allelic association in patients from the Faroe Islands. *Molecular Psychiatry*, **6**, 450-455.

Deloukas, P., Earthrowl, M.E., Grafham, D.V., Rubenfield, M., French, L., Steward, C.A., Sims, S.K., Jones, M.C., Searle, S., Scott, C., Howe, K., Hunt, S.E., Andrews, T.D., Gilbert, J.G.R., Swarbreck, D., Ashurst, J.L., Taylor, A., Battles, J., Bird, C.P., Ainscough, R., Almeida, J.P., Ashwell, R.I.S., Ambrose, K.D., Babbage, A.K., Bagguley, C.L., Bailey, J., Banerjee, R., Bates, K., Beasley, H., Bray-Allen, S., Brown, A.J., Brown, J.Y., Burford, D.C., Burrill, W., Burton, J., Cahill, P., Camire, D., Carter, N.P., Chapman, J.C., Clark, S.Y., Clarke, G., Clee, C.M., Clegg, S., Corby, N., Coulson, A., Dhami, P., Dutta, I., Dunn, M., Faulkner, L., Frankish, A., Frankland, J.A., Garner, P., Garnett, J., Gribble, S., Griffiths, C., Grocock, R., Gustafson, E., Hammond,

S., Harley, J.L., Hart, E., Heath, P.D., Ho, T.P., Hopkins, B., Horne, J., Howden, P.J., Huckle, E., Hynds, C., Johnson, C., Johnson, D., Kana, A., Kay, M., Kimberley, A.M., Kershaw, J.K., Kokkinaki, M., Laird, G.K., Lawlor, S., Lee, H.M., Leongamornlert, D.A., Laird, G., Lloyd, C., Lloyd, D.M., Loveland, J., Lovell, J., McLaren, S., McLay, K.E., McMurray, A., Mashreghi-Mohammadi, M., Matthews, L., Milne, S., Nickerson, T., Nguyen, M., Oveton-Larty, E., Palmer, S.A., Pearce, A.V., Peck, A.I., Pelan, S., Phillimore, B., Porter, K., Rice, C.M., Rogosin, A., Ross, M.T., Sarafidou, T., Sehra, H.K., Shownkeen, R., Skuce, C.D., Smith, M., Standring, L., Sycamore, N., Tester, J., Thorpe, A., Torcasso, W., Tracey, A., Tromans, A., Tsolas, J., Wall, M., Walsh, J., Wang, H., Weinstock, K., West, A.P., Willey, D.L., Whitehead, S.L., Wilming, L., Wray, P.W., Young, L., Chen, Y., Lovering, R.C., Moschonas, N.K., Siebert, R., Fechtel, K., Bentley, D., Durbin, R., Hubbard, T., Doucette-Stamm, L., Beck, S., Smith, D.R. and Rogers, J. (2004) The DNA sequence and comparative analysis of human chromosome 10. *Nature*, **429**, 375-381.

Deninger, P.L. (1989) SINEs: Short interspersed repeated DNA elements in higher eukaryotes. In Berg, D.E. and Howe, M.M. (eds.), *Mobile DNA*. American Society for Microbiology Press, Washington, DC, pp. 619-636.

Devlin, B. and Risch, N. (1995) A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics*, **29**, 311-322.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*, **35**, 41-48.

Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M.L., Haines, G.K. and Barch, D.H. (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics*, **148**, 1269-1284.

Dillehay, T.D. (1997) *Monte Verde: Late Pleistocene Settlement in Chile: The Archaeological Context and Interpretation v. 2 (Smithsonian Series in Archaeological Inquiry)*. Smithsonian Books.

Donaldson, D. (1998) *Psychiatric Disorders with a Biochemical Basis*. The Parthenon Publishing Group Inc. 241.

Drevets, W.C., Frank, E., Price, J.C., Kupfer, D.J., Greer, P.J. and Mathis, C. (2000) Serotonin type-1A receptor imaging in depression. *Nuclear Medicine and Biology*, **27**, 499-507.

Drevets, W.C., Frank, E., Price, J.C., Kupfer, D.J., Holt, D., Greer, P.J., Huang, Y.Y., Gautier, C. and Mathis, C. (1999) PET imaging of serotonin 1A receptor binding in depression. *Biological Psychiatry*, **46**, 1375-1387.

Dreyer, S.D., Zhou, G., Baldini, A., Winterpacht, A., Zabel, B., Cole, W., Johnson, R.L. and Lee, B. (1998) Mutations in LMX1B cause abnormal skeletal patterning and renal dysplasia in nail patella syndrome. *Nature Genetics*, **19**, 47-50.

Duffy, A., Grof, P., Robertson, C. and Alda, M. (2000) The implications of genetic studies of major mood disorders for clinical practice. *Journal of Clinical Psychiatry*, **61**, 630-637.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Balley, J., Barlow, K., Bates, K.N., Beasley, O., Bird, C.P., Blakey, S., Bridgeman, A.M., Buck, D., Burgess, J., Burrill, W.D., Burton, J., Carder, C., Carter, N.P., Chen, Y., Clark, G., Clegg,

S.M., Cobley, V., Cole, C.G., Collier, R.E., Connor, R.E., Conroy, D., Corby, N., Coville, G.J., Cox, A.V., Davis, J., Dawson, E., Dhami, P.D., Dockree, C., Dodsworth, S.J., Durbin, R.M., Ellington, A., Evans, K.L., Fey, J.M., Fleming, K., French, L., Garner, A.A., Gilbert, J.G.R., Goward, M.E., Grafham, D., Griffiths, M.N., Hall, C., Hall, R., Hall-Tamlyn, G., Heathcott, R.W., Ho, S., Holmes, S., Hunt, S.E., Jones, M.C., Kershaw, J., Kimberley, A., King, A., Laird, G.K., Langford, C.F., Leversha, M.A., Lloyd, C., Lloyd, D.M., Martyn, I.D., Mashreghi-Mohammadi, M., Matthews, L., McCann, O.T., McClay, J., McLaren, S., McMurray, A.A., Milne, S.A., Mortimore, B.J., Odell, C.N., Pavitt, R., Pearce, A.V., Pearson, D., Phillimore, B.J., Phillips, S.H., Plumb, R.W., Ramsay, H., Ramsey, Y., Rogers, L., Ross, M.T., Scott, C.E., Sehra, H.K., Skuce, C.D., Smalley, S., Smith, M.L., Soderlund, C., Spragon, L., Steward, C.A., Sulston, J.E., Swann, R.M., Vaudin, M., Wall, M., Wallis, J.M., Whiteley, M.N., Willey, D., Williams, L., Williams, S., Williamson, H., Wilmer, T.E., Wilming, L., Wright, C.L., Hubbard, T., Bentley, D.R., Beck, S., Rogers, J., Minoshima, S., Kawasaki, K., Sasaki, T., Asakawa, S., Kudoh, J., Shintani, A., Shibuya, K., Yoshizaki, Y., Aoki, N., Mitsuyama, S., Chen, F., Chu, L., Crabtree, J., Deschamps, S., Do, A., Do, T., Dorman, A., Fang, F., Fu, Y., Hu, P., Hua, A., Kenton, S., Lai, H., Lao, H.I., Lewis, J., Lewis, S., Lin, S.P., Loh, P., Malaj, E., Nguyen, T., Pan, H., Phan, S., Qi, S., Qian, Y., Ray, L., Ren, Q., Shaull, S., Sloan, D., Song, L., Wang, Q., Wang, Y., Wang, Z., White, J., Willingham, D., Wu, H., Yao, Z., Zhan, M., Zhang, G., Murray, J., Miller, N., Minx, P., Fulton, R., Johnson, D., Bemis, G., Bentley, D., Bradshaw, H., Bourne, S., Cordes, M., Du, Z., Fulton, L., Goela, D., Graves, T., Hawkins, J., Hinds, K., Kemp, K., Latreille, P., Layman, D., Ozersky, P., Rohlfing, T., Scheet, P., Walker, C., Wamsley, A., Wohldmann, P., Pepin, K., Nelson, J., Korf, I., Bedell, J.A., Hillier, L., Mardis, E., Waterston, R., Wilson, R., Emanuel, B.S., Shaikh, T., Kurahashi, H., Saitta, S., Budarf, M.L., McDermid, H.E., Johnson, A., Wong, A.C.C., Morrow, B.E., Edelman, L., Kim, U.J., Shizuya, H., Simon, M.I., Dumanski, J.P., Peyrard, M., Kedra, D., Seroussi, E., Fransson, I., Tapia, I., Bruder, C.E. and O'Brien, K.P. (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489-495.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L. and Jarvela, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, **30**, 233-237.

Engel, L.S., Taioli, E., Pfeiffer, R., Garcia-Closas, M., Marcus, P.M., Lan, Q., Boffetta, P., Vineis, P., Autrup, H., Bell, D.A., Branch, R.A., Brockmoller, J., Daly, A.K., Heckbert, S.R., Kalina, I., Kang, D.H., Katoh, T., Lafuente, A., Lin, H.J., Romkes, M., Taylor, J.A. and Rothman, N. (2002) Pooled analysis and meta-analysis of glutathione S-transferase M1 and bladder cancer: A HuGE review. *American Journal of Epidemiology*, **156**, 95-109.

Escamilla, M.A. (2001) Population isolates: their special value for locating genes for bipolar disorder. *Bipolar Disorders*, **3**, 299-317.

Escamilla, M.A., Freimer, N.B. and Reus, V.I. (1997) The Genetics of Bipolar Disorder and Schizophrenia. In Rosenberg, R., Prusiner, S., DiMauro, S. and Barchi, R. (eds.), *The Molecular and Genetic Basis of Neurological Disease*. Butterworth-Heinemann, Newton, Massachussets, p. 1430.

Escamilla, M.A., McInnes, L.A., Spesny, M., Reus, V.I., Service, S.K., Shimayoshi, N., Tyler, D.J., Silva, S., Molina, J., Gallegos, A., Meza, L., Cruz, M.L., Batki, S., Vinogradov, S., Neylan, T., Nguyen, J.B., Fournier, E., Araya, C.,

Barondes, S.H., Leon, P., Sandkuijl, L.A. and Freimer, N.B. (1999) Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: An initial screen for bipolar disorder loci on chromosome 18. *American Journal of Human Genetics*, **64**, 1670-1678.

Escamilla, M.A., Spesny, M., Reus, V.I., Gallegos, A., Meza, L., Molina, J., Sandkuijl, L.A., Fournier, E., Leon, P.E., Smith, L.B. and Freimer, N.B. (1996) Use of linkage disequilibrium approaches to map genes for bipolar disorder in the Costa Rican population. *American Journal of Medical Genetics*, **67**, 244-253.

Esnault, C., Maestre, J. and Heidmann, T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics*, **24**, 363-367.

Esterling, L.E., Yoshikawa, T., Turner, G., Badner, J.A., Bengel, D., Gershon, E.S., Berrettini, W.H. and Detera-Wadleigh, S.D. (1998) Serotonin transporter (5-HTT) gene and bipolar affective disorder. *American Journal of Medical Genetics*, **81**, 37-40.

Ewald, H., Degn, B., Mors, O. and Krause, T.A. (1998) Significant linkage between bipolar affective disorder and chromosome 12q24. *American Journal of Medical Genetics*, **81**, 540-540.

Ewald, H., Mors, O., Flint, T., Koed, K., Eiberg, H. and Kruse, T.A. (1995) A Possible Locus For Manic-Depressive Illness On Chromosome 16p13. *Psychiatric Genetics*, **5**, 71-81.

Excoffier, L. (2000) Analysis of Population Subdivision. In Balding, D., Bishop, M. and Cannings, C. (eds.), *Handbook of Statistical Genetics*. Wiley and Sons, Ltd.

Excoffier, L. and Schneider, S. (1999) Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 10597-10602.

Excoffier, L. and Slatkin, M. (1998) Incorporating genotypes of relatives into a test of linkage disequilibrium. *American Journal of Human Genetics*, **62**, 171-180.

Fagiolini, A., Kupfer, D.J., Rucci, P., Scott, J.A., Novick, J.M. and Frank, E. (2004) Suicude Attempts and Ideation in Patients with Bipolar I Disorder. *Journal of Clinical Psychiatry*, **65**, 509-514.

Falconer, D. (1981) *Introduction to Quantitative Genetics*. Longman, London, UK.

Falk, C.T. and Rubinstein, P. (1987) Haplotype relative risk: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics*, **51**, 227-233.

Fallin, M.D., Lasseter, V.K., Wolyniec, P.S., McGrath, J.A., Nestadt, G., Valle, D., Liang, K.Y. and Pulver, A.E. (2004) Genomewide linkage scan for bipolar-disorder susceptibility loci among Ashkenazi Jewish families. *American Journal of Human Genetics*, **75**, 204-219.

Feder, J.N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D.A., Basava, A., Dormishian, F., Domingo, R., Ellis, M.C., Fullan, A., Hinton, L.M., Jones, N.L., Kimmel, B.E., Kronmal, G.S., Lauer, P., Lee, V.K., Loeb, D.B., Mapa, F.A., McClelland, E., Meyer, N.C., Mintier, G.A., Moeller, N., Moore, T., Morikang, E., Prass, C.E., Quintana, L., Starnes, S.M., Schatzman, R.C., Brunke, K.J., Drayna, D.T., Risch, N.J., Bacon, B.R. and Wolff, R.K. (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genetics*, **13**, 399-408.

Felsenstein, J. (2004) PHYLIP (Phylogeny Inference Package). Distributed by the author., Department of Genome Sciences, University of Washington, Seattle.

Fields, C., Adams, M.D., White, O. and Venter, J.C. (1994) How Many Genes in the Human Genome. *Nature Genetics*, **7**, 345-346.

Foster, M.W. (2004) Integrating ethics and science in the international HapMap project. *Nature Reviews Genetics*, **5**, 467-475.

Frazer, A. and Hensler, J.G. (1999) Serotonin. In Siegal, G.J., Agranoff, B.W., Albers, R.W., Fisher, S.K. and Uhler, M.D. (eds.), *Basic Neurochemistry: Molecular, Cellular, and Medical Aspects*. Lippincott, Williams & Wilkins, Philadelphia.

Freimer, N.B., Reus, V.I., Escamilla, M., Spesny, M., Smith, L., Service, S., Gallegos, A., Meza, L., Batki, S., Vinogradov, S., Leon, P. and Sandkuijl, L.A. (1996a) An approach to investigating linkage for bipolar disorder using large Costa Rican pedigrees. *American Journal of Medical Genetics*, **67**, 254-263.

Freimer, N.B., Reus, V.I., Escamilla, M.A., McInnes, L.A., Spesny, M., Leon, P., Service, S.K., Smith, L.B., Silva, S., Rojas, E., Gallegos, A., Meza, L., Fournier, E., Baharloo, S., Blankenship, K., Tyler, D.J., Batki, S., Vinogradov, S., Weissenbach, J., Barondes, S.H. and Sandkuijl, L.A. (1996b) Genetic mapping using haplotype, association and linkage methods suggests a locus for severe bipolar disorder (BPI) at 18q22-q23. *Nature Genetics*, **12**, 436-441.

Freimer, N.B., Service, S.K. and Slatkin, M. (1997) Expanding on population studies. *Nature Genetics*, **17**, 371-373.

Fu, Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**, 915-925.

Furlong, R.A., Ho, L., Walsh, C., Rubinsztein, J.S., Jain, S., Paykel, E.S., Easton, D.F. and Rubinsztein, D.C. (1998) Analysis and meta-analysis of two serotonin transporter gene polymorphisms in bipolar and unipolar affective disorders. *American Journal of Medical Genetics*, **81**, 58-63.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225-2229.

Garner, C., McInnes, L.A., Service, S.K., Spesny, M., Fournier, E., Leon, P. and Freimer, N.B. (2001) Linkage analysis of a complex pedigree with severe bipolar disorder, using a Markov chain Monte Carlo method. *American Journal of Human Genetics*, **68**, 1061-1064.

Gelernter, J., Cubells, J.F., Kidd, J.R., Pakstis, A.J. and Kidd, K.K. (1999) Population studies of polymorphisms of the serotonin transporter protein gene. *American Journal of Medical Genetics*, **88**, 61-66.

Gelernter, J., Kranzler, H. and Cubells, J.F. (1997) Serotonin transporter protein (SLC6A4) allele and haplotype frequencies and linkage disequilibria in African- and European- American and Japanese populations and in alcohol-dependent subjects. *Human Genetics*, **101**, 243-246.

Geller, B. and Cook, E.H. (1999) Serotonin transporter gene (HTTLPR) is not in linkage disequilibrium with prepubertal and early adolescent bipolarity. *Biological Psychiatry*, **45**, 1230-1233.

Gershon, E.S., Martinez, M., Goldin, L.R. and Gejman, P.V. (1990) Genetic-Mapping of Common Diseases - the Challenges of Manic- Depressive Illness and Schizophrenia. *Trends in Genetics*, **6**, 282-287.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H.M., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y., Tam, P.K.H., Tsui, L.C., Waye,

M.M.Y., Wong, J.T.F., Zeng, C.Q., Zhang, Q.R., Chee, M.S., Galver, L.M., Kruglyak, S., Murray, S.S., Oliphant, A.R., Montpetit, A., Hudson, T.J., Chagnon, F., Ferretti, V., Leboeuf, M., Phillips, M.S., Verner, A., Kwok, P.Y., Duan, S.H., Lind, D.L., Miller, R.D., Rice, J.P., Saccone, N.L., Taillon-Miller, P., Xiao, M., Nakamura, Y., Sekine, A., Sorimachi, K., Tanaka, T., Tanaka, Y., Tsunoda, T., Yoshino, E., Bentley, D.R., Deloukas, P., Hunt, S., Powell, D., Altshuler, D., Gabriel, S.B., Qiu, R.Z., Ken, A., Dunston, G.M., Kato, K., Niikawa, N., Knoppers, B.M., Foster, M.W., Clayton, E.W., Wang, V.O., Watkin, J., Sodergren, E., Weinstock, G.M., Wilson, R.K., Fulton, L.L., Rogers, J., Birren, B.W., Han, H., Wang, H.G., Godbout, M., Wallenburg, J.C., L'Archeveque, P., Bellemare, G., Todani, K., Fujita, T., Tanaka, S., Holden, A.L., Lai, E.H., Collins, F.S., Brooks, L.D., McEwen, J.E., Guyer, M.S., Jordan, E., Peterson, J.L., Spiegel, J., Sung, L.M., Zacharia, L.F., Kennedy, K., Dunn, M.G., Seabrook, R., Shillito, M., Skene, B., Stewart, J.G., Valle, D.L., Jorde, L.B., Chakravarti, A., Cho, M.K., Duster, T., Jasperse, M., Licinio, J., Long, J.C., Marshall, P.A., Ossorio, P.N., Rotimi, C.N., Royal, C.D.M., Spallone, P., Terry, S.F., Lander, E.S., Nickerson, D.A., Abecasis, G.R., Boehnke, M., Cardon, L.R., Daly, M.J., Douglas, J.A., Hudson, R.R., Kruglyak, L., Nussbaum, R.L., Schaffner, S.F., Sherry, S.T. and Stein, L.D. (2003) The International HapMap Project. *Nature*, **426**, 789-796.

Gilad, Y., Rosenberg, S., Przeworski, M., Lancet, D. and Skorecki, K. (2002) Evidence for positive selection and population structure at the human MAO-A gene. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 862-867.

Ginns, E.I., Ott, J., Egeland, J.A., Allen, C.R., Fann, C.S.J., Pauls, D.L., Weissenbach, J., Carulli, J.P., Falls, K.M., Keith, T.P. and Paul, S.M. (1996) A genome-wide search for chromosomal loci linked to bipolar affective disorder in the Old Order Amish. *Nature Genetics*, **12**, 431-435.

Glaser, B., Kirov, G., Green, E., Craddock, N. and Owen, M.J. (2005) Linkage disequilibrium mapping of bipolar affective disorder at 12q23-q24 provides evidence for association at CUX2 and FLJ32356. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, **132B**, 38-45.

Glatt, C.E., DeYoung, J.A., Delgado, S., Service, S.K., Giacomini, K.M., Edwards, R.H., Risch, N. and Freimer, N.B. (2001) Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. *Nature Genetics*, **27**, 435-438.

Glatt, C.E., Tampilic, M., Christie, C., DeYoung, J. and Freimer, N.B. (2004) Re-screening serotonin receptors for genetic variants identifies population and molecular genetic complexity. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, **124B**, 92-100.

Goldstein, D.B. (2001) Islands of linkage disequilibrium. *Nature Genetics*, **29**, 109-111.

Green, L.D., Derr, J.N. and Knight, A. (2000) mtDNA affinities of the peoples of north-central Mexico. *American Journal of Human Genetics*, **66**, 989-998.

Greenberg, J., Turner, C. and Zegura, S. (1986) The settlement of the Americas: a comparison of the linguistic, dental and genetic evidence. *Current Anthropology*, **27**, 477–497.

Griebel, G. (1995) 5-Hydroxytryptamine-Interacting Drugs in Animal-Models of Anxiety Disorders - More Than 30 Years of Research. *Pharmacology & Therapeutics*, **65**, 319-395.

Gross, C., Zhuang, X.X., Stark, K., Ramboz, S., Oosting, R., Kirby, L., Santarelli, L., Beck, S. and Hen, R. (2002) Serotonin(1A) receptor acts during development to establish normal anxiety-like behaviour in the adult. *Nature*, **416**, 396-400.

Hakak, Y., Walker, J.R., Li, C., Wong, W.H., Davis, K.L., Buxbaum, J.D., Haroutunian, V. and Fienberg, A.A. (2001) Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 4746-4751.

Hall, D., Wijsman, E.M., Roos, J.L., Gogos, J.A. and Karayiorgou, M. (2002) Extended intermarker linkage disequilibrium in the afrikaners. *Genome Research*, **12**, 956-961.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B. and King, M.C. (1990) Linkage of Early-Onset Familial Breast-Cancer to Chromosome- 17q21. *Science*, **250**, 1684-1689.

Hamet, P., Merlo, E., Seda, O., Broeckel, U., Tremblay, J., Kaldunski, M., Gaudet, D., Bouchard, G., Deslauriers, B., Gagnon, F., Antoniol, G., Pausova, Z., Labuda, M., Jomphe, M., Gossard, F., Tremblay, G., Kirova, R., Tonellato, P., Orlov, S.N., Pintos, J., Platko, J., Hudson, T.J., Rioux, J.D., Kotchen, T.A. and Cowley, A.W. (2005) Quantitative founder-effect analysis of French Canadian families identifies specific loci contributing to metabolic phenotypes of hypertension. *American Journal of Human Genetics*, **76**, 815-832.

Hammer, M.F. (1995) A Recent Common Ancestry for Human Y-chromosomes. *Nature*, **378**, 376-378.

Hariri, A.R., Mattay, V.S., Tessitore, A., Kolachana, B., Fera, F., Goldman, D., Egan, M.F. and Weinberger, D.R. (2002) Serotonin transporter genetic variation and the response of the human amygdala. *Science*, **297**, 400-403.

Harpending, H.C., Batzer, M.A., Gurven, M., Jorde, L.B., Rogers, A.R. and Sherry, S.T. (1998) Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 1961-1967.

Hartl, D.L. and Clark, A.G. (1997) *Principles of Population Genetics*. Sinauer Associates Incorporated.

Hatch, F.T., Bodner, A.J., Mazrimas, J.A. and Moore, D.H. (1976) Satellite DNA and Cytogenetic Evolution - DNA Quantity, Satellite DNA and Karyotypic Variations in Kangaroo Rats (Genus-Dipodomys). *Chromosoma*, **58**, 155-168.

Heils, A., Teufel, A., Petri, S., Seemann, M., Bengel, D., Balling, U., Riederer, P. and Lesch, K.P. (1995) Functional promoter and polyadenylation site mapping of the human serotonin (5-HT) transporter gene. *Journal of Neural Transmission-General Section*, **102**, 247-254.

Heils, A., Teufel, A., Petri, S., Stober, G., Riederer, P., Bengel, D. and Lesch, K.P. (1996) Allelic variation of human serotonin transporter gene expression. *Journal of Neurochemistry*, **66**, 2621-2624.

Heisler, L.K., Chu, H.M., Brennan, T.J., Danao, J.A., Bajwa, P., Parsons, L.H. and Tecott, L.H. (1998) Elevated anxiety and antidepressant-like responses in serotonin 5-HT1A receptor mutant mice. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 15049-15054.

Henshilwood, C.S., d'Errico, F., Yates, R., Jacobs, Z., Tribolo, C., Duller, G.A.T., Mercier, N., Sealy, J.C., Valladas, H., Watts, I. and Wintle, A.G. (2002) Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science*, **295**, 1278-1280.

Hey, J. (2004) What's so hot about recombination hotspots? *Plos Biology*, **2**, art. no.-e190.

Hill, W.G. and Weir, B.S. (1994) Maximum-Likelihood-Estimation of Gene Location by Linkage Disequilibrium. *American Journal of Human Genetics*, **54**, 705-714.

Hillier, L.W., Graves, T.A., Fulton, R.S., Fulton, L.A., Pepin, K.H., Minx, P., Wagner-McPherson, C., Layman, D., Wylie, K., Sekhon, M., Becker, M.C., Fewell, G.A., Delehaunty, K.D., Miner, T.L., Nash, W.E., Kremitzki, C., Oddy, L., Du, H., Sun, H., Bradshaw-Cordum, H., Ali, J., Carter, J., Cordes, M., Harris, A., Isak, A., van Brunt, A., Nguyen, C., Du, F.Y., Courtney, L., Kalicki, J., Ozersky, P., Abbott, S., Armstrong, J., Belter, E.A., Caruso, L., Cedroni, M., Cotton, M., Davidson, T., Desai, A., Elliott, G., Erb, T., Fronick, C., Gaige, T., Haakenson, W., Haglund, K., Holmes, A., Harkins, R., Kim, K., Kruchowski, S.S., Strong, C.M., Grewal, N., Goyea, E., Hou, S., Levy, A., Martinka, S., Mead, K., McLellan, M.D., Meyer, R., Maher, J.R., Tomlinson, C., Kohlberg, S.D., Kozlowicz-Reilly, A., Shah, N., Swearengen-Shahid, S., Snider, J., Strong, J.T., Thompson, J., Yoakum, M., Leonard, S., Pearman, C., Trani, L., Radionenko, M., Waligorski, J.E., Wang, C.Y., Rock, S.M., Tin-Wollam, A.M., Maupin, R., Latreille, P., Wendl, M.C., Yang, S.P., Pohl, C., Wallis, J.W., Spieth, J., Bieri, T.A., Berkowicz, N., Nelson, J.O., Osborne, J., Ding, L., Sabo, A., Shotland, Y., Sinha, P., Wohldmann, P.E., Cook, L.L., Hickenbotham, M.T., Eldred, J., Williams, D., Jones, T.A., She, X.W., Ciccarelli, F.D., Izaurralde, E., Taylor, J., Schmutz, J., Myers, R.M., Cox, D.R., Huang, X.Q., McPherson, J.D., Mardis, E.R., Clifton, S.W., Warren, W.C., Chinwalla, A.T., Eddy, S.R., Marra, M.A., Ovcharenko, I., Furey, T.S., Miller, W., Eichler, E.E., Bork, P., Suyama, M., Torrents, D., Waterston, R.H. and Wilson, R.K. (2005) Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature*, **434**, 724-731.

Hirschhorn, J.N. (2005) Genetic approaches to studying common diseases and complex traits. *Pediatric Research*, **57**, 74R-77R.

Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**, 95-108.

Holtkemper, U., Rolf, B., Hohoff, C., Forster, P. and Brinkmannn, B. (2001) Mutation rates at two human Y-chromosomal microsatellite loci using small pool PCR techniques. *Human Molecular Genetics*, **10**, 629-633.

Hoogendoorn, B., Coleman, S.L., Guy, C.A., Smith, K., Bowen, T., Buckland, P.R. and O'Donovan, M.C. (2003) Functional analysis of human promoter polymorphisms. *Human Molecular Genetics*, **12**, 2249-2254.

Huang, Q.Y., Xu, F.H., Shen, H., Deng, H.Y., Liu, Y.J., Liu, Y.Z., Li, J.L., Recker, R.R. and Deng, H.W. (2002) Mutation patterns at dinucleotide microsatellite loci in humans. *American Journal of Human Genetics*, **70**, 625-634.

Hugot, J.P. (2002) Role of NOD2 gene in Crohn's disease. *Gastroenterologie Clinique Et Biologique*, **26**, 13-15.

Hui, J.Y., Stangl, K., Lane, W.S. and Bindereif, A. (2003) HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nature Structural Biology*, **10**, 33-37.

Humphray, S.J., Oliver, K., Hunt, A.R., Plumb, R.W., Loveland, J.E., Howe, K.L., Andrews, T.D., Searle, S., Hunt, S.E., Scott, C.E., Jones, M.C., Ainscough, R., Almeida, J.P., Ambrose, K.D., Ashwell, R.I.S., Babbage, A.K., Babbage, S., Bagguley, C.L., Bailey, J., Banerjee, R., Barker, D.J., Barlow, K.F., Bates, K.,

Beasley, H., Beasley, O., Bird, C.P., Bray-Allen, S., Brown, A.J., Brown, J.Y., Burford, D., Burrill, W., Burton, J., Carder, C., Carter, N.P., Chapman, J.C., Chen, Y., Clarke, G., Clark, S.Y., Clee, C.M., Clegg, S., Collier, R.E., Corby, N., Crosier, M., Cummings, A.T., Davies, J., Dhami, P., Dunn, M., Dutta, I., Dyer, L.W., Earthrowl, M.E., Faulkner, L., Fleming, C.J., Frankish, A., Frankland, J.A., French, L., Fricker, D.G., Garner, P., Garnett, J., Ghori, J., Gilbert, J.G.R., Glison, C., Grafham, D.V., Gribble, S., Griffiths, C., Jones, S.G., Grocock, R., Guy, J., Hall, R.E., Hammond, S., Harley, J.L., Harrison, E.S.I., Hart, E.A., Heath, P.D., Henderson, C.D., Hopkins, B.L., Howard, P.J., Howden, P.J., Huckle, E., Johnson, C., Johnson, D., Joy, A.A., Kay, M., Keenan, S., Kershaw, J.K., Kimberley, A.M., King, A., Knights, A., Laird, G.K., Langford, C., Lawlor, S., Leongamornlert, D.A., Leversha, M., Lloyd, C., Lloyd, D.M., Lovell, J., Martin, S., Mashreghi-Mohammadi, M., Matthews, L., McLaren, S., McLay, K.E., McMurray, A., Milne, S., Nickerson, T., Nisbett, J., Nordsiek, G., Pearce, A.V., Peck, A.I., Porter, K.M., Pandian, R., Pelan, S., Phillimore, B., Povey, S., Ramsey, Y., Rand, V., Scharfe, M., Sehra, H.K., Shownkeen, R., Sims, S.K., Skuce, C.D., Smith, M., Steward, C.A., Swarbreck, D., Sycamore, N., Tester, J., Thorpe, A., Tracey, A., Tromans, A., Thomas, D.W., Wall, M., Wallis, J.M., West, A.P., Whitehead, S.L., Willey, D.L., Williams, S.A., Wilming, L., Wray, P.W., Young, L., Ashurst, J.L., Coulson, A., Blocker, H., Durbin, R., Sulston, J.E., Hubbard, T., Jackson, M.J., Bentley, D.R., Beck, S., Rogers, J. and Dunham, I. (2004) DNA sequence and analysis of human chromosome 9. *Nature*, **429**, 369-374.

Hurles, M.E., Willey, D., Matthews, L. and Hussain, S.S. (2004) Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. *Genome Biology*, **5**, art. no.-R55.

Hutchison, C.A.I., Hardies, S.C., Loeb, D.D., Shehee, W.R. and Edgell, M.H. (1989) LINES and related retroposons: Long interspersed repeated sequences in the eukaryotic genome. In Berg, D.E. and Howe, M.M. (eds.), *Mobile DNA*. American Society for Microbiology Press, Washington, DC, pp. 593-617.

Jablensky, A. (2000) Epidemiology of schizophrenia: the global burden of disease and disability. *European Archives of Psychiatry and Clinical Neuroscience*, **250**, 274-285.

Jeffreys, A., Barber, R., Bois, P., Buard, J., Dubrova, Y.E., Grant, G., Hollies, C.R.H., May, C.A., Neumann, R., Panayi, M., Ritchie, A.E., Shone, A.C., Signer, E., Stead, J.D.H. and Tamaki, K. (1999) Human minisatellites, repeat DNA instability and meiotic recombination. *Electrophoresis*, **20**, 1665-1675.

Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**, 217-222.

Jeffreys, A.J., Neil, D.L. and Neumann, R. (1998) Repeat instability at human minisatellites arising from meiotic recombination. *Embo Journal*, **17**, 4147-4157.

Jeffreys, A.J., Neumann, R., Panayi, M., Myers, S. and Donnelly, P. (2005) Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics*, **37**, 601-606.

Jobling, M.A., Hurles, M.E. and Tyler-Smith, C. (2004) *Human evolutionary genetics : origins, peoples & disease*. Garland Science, New York.

Johnson, G.C.L., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., Twells, R.C.J., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S.C.L., Clayton, D.G. and Todd, J.A. (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics*, **29**, 233-237.

Jones, S., Martin, R.D. and (editors), D.P. (1994) *The Cambridge Encyclopedia of Human Evolution*. Cambridge University Press, Cambridge.

Jorde, L.B., Rogers, A.R., Bamshad, M., Watkins, W.S., Krakowiak, P., Sung, S., Kere, J. and Harpending, H.C. (1997) Microsatellite diversity and the demographic history of modern humans. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 3100-3103.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T. and Batzer, M.A. (2000) The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. *American Journal of Human Genetics*, **66**, 979-988.

Kaessmann, H., Wiebe, V. and Paabo, S. (1999) Extensive nuclear DNA sequence diversity among chimpanzees. *Science*, **286**, 1159-1162.

Kaessmann, H., Wiebe, V., Weiss, G. and Paabo, S. (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genetics*, **27**, 155-156.

Kauppi, L., Jeffreys, A.J. and Keeney, S. (2004) Where the crossovers are: Recombination distributions in mammals. *Nature Reviews Genetics*, **5**, 413-424.

Kawanishi, Y., Harada, S., Tachikawa, H., Okubo, T. and Shiraishi, H. (1998) Novel mutations in the promoter and coding region of the human 5-HT1A receptor gene and association analysis in schizophrenia. *American Journal of Medical Genetics*, **81**, 434-439.

Kayser, M., Brauer, S. and Stoneking, M. (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*, **20**, 893-900.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff, P., Stoneking, M. and Sajantila, A. (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *American Journal of Human Genetics*, **66**, 1580-1588.

Kayser, M. and Sajantila, A. (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Science International*, **118**, 116-121.

Kelsoe, J.R., Sadovnick, A.D., Kristbjarnarson, H., Bergesch, P., MroczkowskiParker, Z., Drennan, M., Rapaport, M.H., Flodman, P., Spence, M.A. and Remick, R.A. (1996) Possible locus of bipolar disorder near the dopamine transporter on chromosome 5. *American Journal of Medical Genetics*, **67**, 533-540.

Kendler, K.S., Neale, M., Kessler, R., Heath, A. and Eaves, L. (1993) A Twin Study of Recent Life Events and Difficulties. *Archives of General Psychiatry*, **50**, 789-796.

Kendler, K.S., Pedersen, N.L., Neale, M.C. and Mathe, A.A. (1995a) A Pilot Swedish Twin Study of Affective-Illness Including Hospital-Ascertained and

Population-Ascertained Subsamples - Results of Model-Fitting. *Behavior Genetics*, **25**, 217-232.

Kendler, K.S., Walters, E.E., Neale, M.C., Kessler, R.C., Heath, A.C. and Eaves, L.J. (1995b) The Structure of the Genetic and Environmental Risk-Factors for 6 Major Psychiatric-Disorders in Women - Phobia, Generalized Anxiety Disorder, Panic Disorder, Bulimia, Major Depression, and Alcoholism. *Archives of General Psychiatry*, **52**, 374-383.

Kessler, R.C., McGonagle, K.A., Zhao, S.Y., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U. and Kendler, K.S. (1994) Lifetime and 12-Month Prevalence of Dsm-Iii-R Psychiatric- Disorders in the United-States - Results From the National- Comorbidity-Survey. *Archives of General Psychiatry*, **51**, 8-19.

Kimmel, M., Chakraborty, R., King, J.P., Bamshad, M., Watkins, W.S. and Jorde, L.B. (1998) Signatures of population expansion in microsatellite repeat data. *Genetics*, **148**, 1921-1930.

Kimura, M. (1968) Evolutionary Rate at Molecular Level. *Nature*, **217**, 624-&.

King, M.C. and Wilson, A.C. (1975) Evolution at 2 Levels in Humans and Chimpanzees. *Science*, **188**, 107-116.

Kirch, P.V. (1999) *The Lapita Peoples*. Blackwell, Oxford, U.K.

Kirkwood, B.R. (1988) *Essentials of Medical Statistics*. Blackwell Scientific Publications, Oxford.

Kirov, G., Rees, M., Jones, I., MacCandless, F., Owen, M.J. and Craddock, N. (1999) Bipolar disorder and the serotonin transporter gene: a family- based association study. *Psychological Medicine*, **29**, 1249-1254.

Kloor, M., Becker, C., Benner, A., Woerner, S.M., Gebert, J., Ferrone, S. and Doeberitz, M.V. (2005) Immunoselective pressure and human leukocyte antigen class I antigen machinery defects in microsatellite unstable colorectal cancers. *Cancer Research*, **65**, 6418-6424.

Knowles, J.A., Rao, P.A., Cox-Matise, T., Loth, J.E., de Jesus, G.M., Levine, L., Das, K., Penchaszadeh, G.K., Alexander, J.R., Lerer, B., Endicott, J., Ott, J., Gilliam, T.C. and Baron, M. (1998) No evidence for significant linkage between bipolar affective disorder and chromosome 18 pericentromeric markers in a large series of multiplex extended pedigrees. *American Journal of Human Genetics*, **62**, 916-924.

Kobilka, B.K., Frielle, T., Collins, S., Yangfeng, T., Kobilka, T.S., Francke, U., Lefkowitz, R.J. and Caron, M.G. (1987) An Intronless Gene Encoding a Potential Member of the Family of Receptors Coupled to Guanine-Nucleotide Regulatory Proteins. *Nature*, **329**, 75-79.

Koonin, E.V.a.G., M. Y. (2003) *Sequence - evolution - function : computational approaches in comparative genomics*. Kluwer Academic Publishers.

Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**, 139-144.

Kruglyak, L., Daly, M.J., ReeveDaly, M.P. and Lander, E.S. (1996) Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347-1363.

Kumar, S. and Subramanian, S. (2002) Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 803-808.

Kwok, P.Y., Deng, Q., Zakeri, H., Taylor, S.L. and Nickerson, D.A. (1996) Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics*, **31**, 123-126.

Laan, M. and Paabo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nature Genetics*, **17**, 435-438.

Lander, E.S. (1996) The new genomics: Global views of biology. *Science*, **274**, 536-539.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H.M., Yu, J., Wang, J., Huang, G.Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S.Z., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W.H., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J.R., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A.,

Patrinos, A. and Morgan, M.J. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Lasky-Su, J.A., Faraone, S.V., Glatt, S.J. and Tsuang, M.T. (2005) Meta-analysis of the association between two polymorphisms in the serotonin transporter gene and affective disorders. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, **133B**, 110-115.

Lazzeroni, L.C. and Lange, K. (1998) A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity*, **48**, 67-81.

Leakey, L.S.B., Napier, J.R. and Tobias, P.V. (1964) New Species of Genus Homo from Olduvai Gorge. *Nature*, **202**, 7-&.

Lell, J.T., Sukernik, R.I., Starikovskaya, Y.B., Su, B., Jin, L., Schurr, T.G., Underhill, P.A. and Wallace, D.C. (2002) The dual origin and Siberian affinities of native American Y chromosomes. *American Journal of Human Genetics*, **70**, 192-206.

Lendon, C.L., Martinez, A., Behrens, I.M., Kosik, K.S., Madrigal, L., Norton, J., Neuman, R., Myers, A., Busfield, F., Wragg, M., Arcos, M., Viana, J.C.A., Ossa, J., Ruiz, A., Goate, A.M. and Lopera, F. (1997) E280A PS-1 mutation causes Alzheimer's disease but age of onset is not modified by ApoE alleles. *Human Mutation*, **10**, 186-195.

Lerer, B., Macciardi, F., Segman, R.H., Adolfsson, R., Blackwood, D., Blairy, S., Del Favero, J., Dikeos, D.G., Kaneva, R., Lilli, R., Massat, I., Milanova, V., Muir, W., Noethen, M., Oruc, L., Petrova, T., Papadimitriou, G.N., Rietschel, M., Serretti, A., Souery, D., Van Gestel, S., Van Broeckhoven, C. and Mendlewicz, J. (2001) Variability of 5-HT2C receptor cys23ser polymorphism among European populations and vulnerability to affective disorder. *Molecular Psychiatry*, **6**, 579-585.

Lesch, K.P. (2001a) Serotonergic gene expression and depression: implications for developing novel antidepressants. *Journal of Affective Disorders*, **62**, 57-76.

Lesch, K.P. (2001b) Variation of serotonergic gene expression: neurodevelopment and the complexity of response to psychopharmacologic drugs. *European Neuropsychopharmacology*, **11**, 457-474.

Lesch, K.P., Balling, U., Gross, J., Strauss, K., Wolozin, B.L., Murphy, D.L. and Riederer, P. (1994) Organization of the Human Serotonin Transporter Gene. *Journal of Neural Transmission-General Section*, **95**, 157-162.

Lesch, K.P., Bengel, D., Heils, A., Sabol, S.Z., Greenberg, B.D., Petri, S., Benjamin, J., Muller, C.R., Hamer, D.H. and Murphy, D.L. (1996) Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science*, **274**, 1527-1531.

Lesch, K.P. and Mossner, R. (1998) Genetically driven variation in serotonin uptake: Is there a link to affective spectrum, neurodevelopmental, and neurodegenerative disorders? *Biological Psychiatry*, **44**, 179-192.

Lewontin, R.C. (1964) Interaction of Selection + Linkage .I. General Considerations - Heterotic Models. *Genetics*, **49**, 49-&.

Lewontin, R.C. and Krakauer, J. (1973) Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms. *Genetics*, **74**, 175-195.

Li, T., Stefansson, H., Gudfinnsson, E., Cai, G., Liu, X., Murray, R.M., Steinthorsdottir, V., Januel, D., Gudnadottir, V.G., Petursson, H., Ingason, A., Gulcher, J.R., Stefansson, K. and Collier, D.A. (2004) Identification of a novel neuregulin 1 at-risk haplotype in Han schizophrenia Chinese patients, but no

association with the Icelandic/Scottish risk haplotype. *Molecular Psychiatry*, **9**, 698-704.

Li, T., Vallada, H., Curtis, D., Arranz, M., Xu, K., Cai, G.Q., Deng, H., Liu, J., Murray, R., Liu, X.H. and Collier, D.A. (1997) Catechol-O-methyltransferase Val158Met polymorphism: frequency analysis in Han Chinese subjects and allelic association of the low activity allele with bipolar affective disorder. *Pharmacogenetics*, **7**, 349-353.

Li, W.-H. (1997) *Fundamentals of Molecular Evolution*. Sinauer Associates Incorporated, Sunderland, MA, U.S.A.

Lieberman, D.E., McBratney, B.M. and Krovitz, G. (2002) The evolution and development of cranial form in Homo sapiens. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 1134-1139.

Lim, L.C.C., Powell, J., Sham, P., Castle, D., Hunt, N., Murray, R. and Gill, M. (1995) Evidence For a Genetic Association Between Alleles of Monoamine-Oxidase a Gene and Bipolar Affective-Disorder. *American Journal of Medical Genetics*, **60**, 325-331.

Little, K.Y., Krolewski, D.M., Zhang, L. and Cassin, B.J. (2003) Loss of striatal vesicular monoamine transporter protein (VMAT2) in human cocaine users. *American Journal of Psychiatry*, **160**, 47-55.

Liu, K. and Muse, S. PowerMarker: new genetic data analysis software.

Liu, K.J. and Muse, S.V. (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, **21**, 2128-2129.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, **33**, 177-182.

Lucki, I. (1998) The spectrum of behaviors influenced by serotonin. *Biological Psychiatry*, **44**, 151-162.

Lupski, J.R. (2004) Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. *Genome Biology*, **5**, art. no.-242.

Lyonnet, S., Bolino, A., Pelet, A., Abel, L., Nihoulfekete, C., Briard, M.L., Moksiu, V., Kaariainen, H., Martucciello, G., Lerone, M., Puliti, A., Luo, Y., Weissenbach, J., Devoto, M., Munnich, A. and Romeo, G. (1993) A Gene for Hirschsprung Disease Maps to the Proximal Long Arm of Chromosome-10. *Nature Genetics*, **4**, 346-350.

Macdonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., Groot, N., Macfarlane, H., Jenkins, B., Anderson, M.A., Wexler, N.S., Gusella, J.F., Bates, G.P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.M., Lehrach, H., Buckler, A.J., Church, D., Doucettestamm, L., Odonovan, M.C., Ribaramirez, L., Shah, M., Stanton, V.P., Strobel, S.A., Draths, K.M., Wales, J.L., Dervan, P., Housman, D.E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J.J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F.S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D. and Harper, P.S. (1993) A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntingtons-Disease Chromosomes. *Cell*, **72**, 971-983.

Mackay, T.F.C. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics*, **35**, 303-339.

MacKenzie, A. and Quinn, J. (1999) A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential enhancer-like properties in the mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 15251-15255.

MacKinnon, D.F., Jamison, K.R. and DePaulo, J.R. (1997) Genetics of manic depressive illness. *Annual Review of Neuroscience*, **20**, 355-373.

Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) Regulation of Inducible and Tissue-Specific Gene-Expression. *Science*, **236**, 1237-1245.

Marth, G., Schuler, G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Wheelan, S., Baker, J., Ward, M., Kholodov, M., Phan, L., Czabarka, E., Murvai, J., Cutler, D., Wooding, S., Rogers, A., Chakravarti, A., Harpending, H.C., Kwok, P.Y. and Sherry, S.T. (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 376-381.

Massat, I., Souery, D., Del-Favero, J., Van Gestel, S., Serretti, A., Macciardi, F., Smeraldi, E., Kaneva, R., Adolfsson, R., Nylander, P.O., Blackwood, D., Muir, W., Papadimitriou, G.N., Dikeos, D., Oruc, L., Segman, R.H., Ivezic, S., Aschauer, H., Ackenheil, M., Fuchshuber, S., Dam, H., Jakovljevic, M., Peltonen, L., Hilger, C., Hentges, F., Staner, L., Milanova, V., Jazin, E., Lerer, B., Van Broeckhoven, C. and Mendlewicz, J. (2002) Positive association of dopamine D2 receptor polymorphism with bipolar affective disorder in a European multicenter association study of affective disorders. *American Journal of Medical Genetics*, **114**, 177-185.

McCauley, J.L., Olson, L.M., Dowd, M., Amin, T., Steele, A., Blakely, R.D., Folstein, S.E., Haines, J.L. and Sutcliffe, J.S. (2004) Linkage and association analysis at the serotonin transporter (SLC6A4) locus in a rigid-compulsive subset of autism. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, **127B**, 104-112.

McGuffin, P. and Katz, R. (1989) The Genetics of Depression and Manic-Depressive Disorder. *British Journal of Psychiatry*, **155**, 294-304.

McInnes, L.A., Escamilla, M.A., Service, S.K., Reus, V.I., Leon, P., Silva, S., Rojas, E., Spesny, M., Baharloo, S., Blankenship, K., Peterson, A., Tyler, D., Shimayoshi, N., Tobey, C., Batki, S., Vinogradov, S., Meza, L., Gallegos, A., Fournier, E., Smith, L.B., Barondes, S.H., Sandkuijl, L.A. and Freimer, N.B. (1996) A complete genome screen for genes predisposing to severe bipolar disorder in two Costa Rican pedigrees. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13060-13065.

McInnes, L.A., Service, S.K., Reus, V.I., Barnes, G., Charlat, O., Jawahar, S., Lewitzky, S., Yang, Q., Duong, Q.Y., Spesny, M., Araya, C., Araya, X., Gallegos, A., Meza, L., Molina, J., Ramirez, R., Mendez, R., Silva, S., Fournier, E., Batki, S.L., Mathews, C.A., Neylan, T., Glatt, C.E., Escamilla, M.A., Luo, D., Gajiwala, P., Song, T., Crook, S., Nguyen, J.B., Roche, E., Meyer, J.M., Leon, P., Sandkuijl, L.A., Freimer, N.B. and Chen, H. (2001) Fine-scale mapping of a locus for severe bipolar mood disorder on chromosome 18p11.3 in the Costa Rican population. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 11485-11490.

McMahon, F.J., Simpson, S.G., McInnis, M.G., Badner, J.A., MacKinnon, D.F. and DePaulo, J.R. (2001) Linkage of bipolar disorder to chromosome 18q and the validity of bipolar II disorder. *Archives of General Psychiatry*, **58**, 1025-1031.

McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581-584.

Medicine, N.L.o. (2003) Health Services/Technology Assessment Text (HSTAT). National Library of Medicine (US), Bethesda (MD).

Meloni, R., Leboyer, M., Bellivier, F., Barbe, B., Samolyk, D., Allilaire, J.F. and Mallet, J. (1995) Association of Manic-Depressive Illness With Tyrosine-Hydroxylase Microsatellite Marker. *Lancet*, **345**, 932-932.

Meltzer, D.J. (1995) Clocking the First Americans. *Annual Review of Anthropology*, **24**, 21-45.

Meltzer, D.J. (1997) Anthropology - Monte Verde and the Pleistocene peopling of the Americas. *Science*, **276**, 754-755.

Merriwether, D.A. and Ferrell, R.E. (1996) The four founding lineage hypothesis for the New World: A critical reevaluation. *Molecular Phylogenetics and Evolution*, **5**, 241-246.

Merriwether, D.A., Friedlaender, J.S., Mediavilla, J., Mgone, C., Gentz, F. and Ferrell, R.E. (1999) Mitochondrial DNA variation is an indicator of austronesian influence in Island Melanesia. *American Journal of Physical Anthropology*, **110**, 243-270.

Merriwether, D.A., Hall, W.W., Vahlne, A. and Ferrell, R.E. (1996) mtDNA variation indicates Mongolia may have been the source for the founding population for the New World. *American Journal of Human Genetics*, **59**, 204-212.

Merriwether, D.A., Huston, S., Iyengar, S., Hamman, R., Norris, J.M., Shetterly, S.M., Kamboh, M.I. and Ferrell, R.E. (1997) Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating. *American Journal of Physical Anthropology*, **102**, 153-159.

Mesa, N.R., Mondragon, M.C., Soto, I.D., Parra, M.V., Duque, C., Ortiz-Barrientos, D., Garcia, L.F., Velez, I.D., Bravo, M.L., Munera, J.G., Bedoya, G., Bortolini, M.C. and Ruiz-Linares, A. (2000) Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: Pre- and post-Columbian patterns of gene flow in South America. *American Journal of Human Genetics*, **67**, 1277-1286.

Miller, R.D. and Kwok, P.Y. (2001) The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Human Molecular Genetics*, **10**, 2195-2198.

Morissette, J., Villeneuve, A., Bordeleau, L., Rochette, D., Laberge, C., Gagne, B., Laprise, C., Bouchard, G., Plante, M., Gobeil, L., Shink, E., Weissenbach, J. and Barden, N. (1999) Genome-wide search for linkage of bipolar affective disorders in a very large pedigree derived from a homogeneous population in Quebec points to a locus of major effect on chromosome 12q23-q24. *American Journal of Medical Genetics*, **88**, 567-587.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743-747.

Mountain, J.L. and Cavalli-Sforza, L.L. (1994) Inference of Human-Evolution through Cladistic-Analysis of Nuclear-DNA Restriction Polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 6515-6519.

Muller-Oerlinghausen, B., Berghofer, A. and Bauer, M. (2002) Bipolar disorder. *Lancet*, **359**, 241-247.

Mynett-Johnson, L., Kealey, C., Claffey, E., Curtis, D., Bouchier-Hayes, L., Powell, C. and McKeon, P. (2000) Multimarkerhaplotypes within the serotonin transporter gene suggest evidence of an association with bipolar disorder. *American Journal of Medical Genetics*, **96**, 845-849.

Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297-304.

Nakahori, Y., Mitani, K., Yamada, M. and Nakagome, Y. (1986) A Human Y-Chromosome Specific Repeated DNA Family (Dyz1) Consists of a Tandem Array of Pentanucleotides. *Nucleic Acids Research*, **14**, 7569-7580.

Nakamura, M., Ueno, S., Sano, A. and Tanabe, H. (2000) The human serotonin transporter gene linked polymorphism (5- HTTLPR) shows ten novel allelic variants. *Molecular Psychiatry*, **5**, 32-38.

Nakhai, B., Nielsen, D.A., Linnoila, M. and Goldman, D. (1995) 2 Naturally-Occurring Amino-Acid Substitutions in the Human 5- Ht1a Receptor - Glycine-22 to Serine-22 and Isoleucine-28 to Valine-28. *Biochemical and Biophysical Research Communications*, **210**, 530-536.

Neale, B.M. and Sham, P.C. (2004) The future of association studies: Gene-based analysis and replication. *American Journal of Human Genetics*, **75**, 353-362.

Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University press, New York, NY, USA.

Neumeister, A., Bain, E., Nugent, A.C., Carson, R.E., Bonne, O., Luckenbaugh, D.A., Eckelman, W., Herscovitch, P., Charney, D.S. and Drevets, W.C. (2004) Reduced serotonin type 1(A) receptor binding in panic disorder. *Journal of Neuroscience*, **24**, 589-591.

Nickerson, D.A., Tobe, V.O. and Taylor, S.L. (1997) PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745-2751.

Nikitina, T.V. and Nazarenko, S.A. (2004) Human microsatellites: Mutation and evolution. *Russian Journal of Genetics*, **40**, 1065-1079.

Nurnberger, J.I., Blehar, M.C., Kaufmann, C.A., Yorkcooler, C., Simpson, S.G., Harkavyfriedman, J., Severe, J.B., Malaspina, D., Reich, T., Miller, M., Bowman, E.S., Depaulo, J.R., Cloninger, C.R., Robinson, G., Modlin, S., Gershon, E.S., Maxwell, E., Guroff, J.J., Kirch, D., Wynne, D., Berg, K., Tsuang, M.T., Faraone, S.V., Pepple, J.R. and Ritz, A.L. (1994) Diagnostic Interview for Genetic-Studies - Rationale, Unique Features, and Training. *Archives of General Psychiatry*, **51**, 849-859.

Ogilvie, A.D., Battersby, S., Bubb, V.J., Fink, G., Harmar, A.J., Goodwin, G.M. and Smith, C.A.D. (1996) Polymorphism in serotonin transporter gene associated with susceptibility to major depression. *Lancet*, **347**, 731-733.

Oleksiak, M.F., Churchill, G.A. and Crawford, D.L. (2002) Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261-266.

Ophoff, R.A., Escamilla, M.A., Service, S.K., Spesny, M., Meshi, D.B., Poon, W., Molina, J., Fournier, E., Gallegos, A., Mathews, C., Neylan, T., Batki, S.L., Roche, E., Ramirez, M., Silva, S., De Mille, M.C., Dong, P., Leon, P.E., Reus, V.I., Sandkuijl, L.A. and Freimer, N.B. (2002) Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate. *American Journal of Human Genetics*, **71**, 565-574.

Ospina-Duque, J., Duque, C., Carvajal-Carmona, L., Ortiz-Barrientos, D., Soto, I., Pineda, N., Cuartas, M., Calle, J., Lopez, C., Ochoa, L., Garcia, J., Gomez, J., Agudelo, A., Lozano, M., Montoya, G., Ospina, A., Lopez, M., Gallo, A., Miranda, A., Serna, L., Montoya, P., Palacio, C., Bedoya, G., McCarthy, M., Reus, V., Freimer, N. and Ruiz-Linares, A. (2000) An association study of bipolar mood disorder (type I) with the 5-HTTLPR serotonin transporter polymorphism in a human population isolate from Colombia. *Neuroscience Letters*, **292**, 199-202.

Owen, M., Cardno, A.G. and O'Donovan, M.C. (2000) Psychiatric genetics: back to the future. *Molecular Psychiatry*, **5**, 22-31.

Owens, M.J. and Nemeroff, C.B. (1994) Role of Serotonin in the Pathophysiology of Depression - Focus on the Serotonin Transporter. *Clinical Chemistry*, **40**, 288-295.

Ozaki, N., Goldman, D., Kaye, W.H., Plotnicov, K., Greenberg, B.D., Lappalainen, J., Rudnick, G. and Murphy, D.L. (2003) Serotonin transporter missense mutation associated with a complex neuropsychiatric phenotype. *Molecular Psychiatry*, **8**, 933-936.

Parks, C.L., Robinson, P.S., Sibille, E., Shenk, T. and Toth, M. (1998) Increased anxiety of mice lacking the serotonin(1A) receptor. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 10734-10739.

Parks, C.L. and Shenk, T. (1996) The serotonin 1a receptor gene contains a TATA-less promoter that responds to MAZ and Sp1. *Journal of Biological Chemistry*, **271**, 4417-4430.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D., Daly, M.J. and Reich, D. (2004) Methods for high-density admixture mapping of disease genes. *American Journal of Human Genetics*, **74**, 979-1000.

Peltonen, L. (2000) Positional cloning of disease genes: Advantages of genetic isolates. *Human Heredity*, **50**, 66-75.

Peter, D., Finn, J.P., Klisak, I., Liu, Y.J., Kojis, T., Heinzmann, C., Roghani, A., Sparkes, R.S. and Edwards, R.H. (1993) Chromosomal Localization of the Human Vesicular Amine Transporter Genes. *Genomics*, **18**, 720-723.

Petrov, D.A. (2001) Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, **17**, 23-28.

Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E. and Shriver, M.D. (2001) Population structure in admired populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. *American Journal of Human Genetics*, **68**, 198-207.

Piccardi, M.P., Ardau, R., Chillotti, C., Deleuze, J.F., Mallet, J., Meloni, R., Oi, A., Severino, G., Congiu, D., Bayorek, M. and Del Zompo, M. (2002) Manic-depressive illness: an association study with the inositol polyphosphate 1-phosphatase and serotonin transporter genes. *Psychiatric Genetics*, **12**, 23-27.

Pineda-Trujillo, N.L., Carvajal-Carmona, L.G., Buritica, O., Moreno, S., Uribe, C., Pineda, D., Toro, M., Garcia, F., Arias, W., Bedoya, G., Lopera, F. and Ruiz-Lineares, A. (2001) A novel Cys212Tyr founder mutation in parkin and allelic heterogeneity of juvenile Parkinsonism in a population from North West Colombia. *Neuroscience Letters*, **298**, 87-90.

Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, **69**, 124-137.

Pritchard, J.K. and Cox, N.J. (2002) The allelic architecture of human disease genes: common disease - common variant ... or not? *Human Molecular Genetics*, **11**, 2417-2423.

Puffenberger, E.G., Kauffman, E.R., Bolk, S., Matise, T.C., Washington, S.S., Angrist, M., Weissenbach, J., Garver, K.L., Mascari, M., Ladda, R., Slaugenhaupt, S.A. and Chakravarti, A. (1994) Identity-by-Descent and Association Mapping of a Recessive Gene for Hirschsprung Disease on Human-Chromosome 13q22. *Human Molecular Genetics*, **3**, 1217-1225.

Quintana-Murci, L., Semino, O., Poloni, E.S., Liu, A., Van Gijn, M., Passarino, G., Brega, A., Nasidze, I.S., Maccioni, L., Cossu, G., Al-Zahery, N., Kidd, J.R., Kidd, K.K. and Santachiara-Benerecetti, A.S. (1999) Y-chromosome specific YCAII, DYS19 and YAP polymorphisms in human populations: a comparative study. *Annals of Human Genetics*, **63**, 153-166.

Ramamoorthy, S., Bauman, A.L., Moore, K.R., Han, H., Yangfeng, T., Chang, A.S., Ganapathy, V. and Blakely, R.D. (1993) Antidepressant-Sensitive and Cocaine-Sensitive Human Serotonin Transporter - Molecular-Cloning, Expression, and Chromosomal Localization. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 2542-2546.

Ramboz, S., Oosting, R., Amara, D.A., Kung, H.F., Blier, P., Mendelsohn, M., Mann, J.J., Brunner, D. and Hen, R. (1998) Serotonin receptor 1A knockout: An animal model of anxiety- related disorder. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14476-14481.

Rees, M., Norton, N., Jones, I., McCandless, F., Scourfield, J., Holmans, P., Moorhead, S., Feldman, E., Sadler, S., Cole, T., Redman, K., Farmer, A., McGuffin, P., Owen, M.J. and Craddock, N. (1997) Association studies of bipolar disorder at the human serotonin transporter gene (hSERT; 5HTT). *Molecular Psychiatry*, **2**, 398-402.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199-204.

Reich, D.E. and Goldstein, D.B. (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 8119-8123.

Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends in Genetics*, **17**, 502-510.

Reich, T., Clayton, P.J. and Winokur, G. (1969) Family History Studies .V. Genetics of Mania. *American Journal of Psychiatry*, **125**, 1358-&.

Reus, V.I. and Freimer, N.B. (1997) Understanding the genetic basis of mood disorders: Where do we stand? *American Journal of Human Genetics*, **60**, 1283-1288.

Riordan, J.R., Rommens, J.M., Kerem, B.S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L., Drumm, M.L., Iannuzzi, M.C., Collins, F.S. and Tsui, L.C. (1989) Identification of the Cystic-Fibrosis Gene - Cloning and Characterization of Complementary-DNA. *Science*, **245**, 1066-1072.

Risch, N., Burchard, E., Ziv, E. and Tang, H. (2002) Categorization of humans in biomedical research: genes, race and disease. *Genome Biology*, **3** (7), 2007.1-2007.12.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516-1517.

Robertson, K.D. and Wolffe, A.P. (2000) DNA methylation in health and disease. *Nature Reviews Genetics*, **1**, 11-19.

Robins, L.N., Helzer, J.E., Weissman, M.M., Orvaschel, H., Gruenberg, E., Burke, J.D. and Regier, D.A. (1984) Lifetime Prevalence of Specific Psychiatric-Disorders in 3 Sites. *Archives of General Psychiatry*, **41**, 949-958.

Rogers, A.R. and Harpending, H. (1992) Population-Growth Makes Waves in the Distribution of Pairwise Genetic-Differences. *Molecular Biology and Evolution*, **9**, 552-569.

Romualdi, C., Balding, D., Nasidze, I.S., Risch, G., Robichaux, M., Sherry, S.T., Stoneking, M., Batzer, M.A. and Barbujani, G. (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Research*, **12**, 602-612.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381-2385.

Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P., Frankish, A., Lovell, F.L., Howe, K.L., Ashurst, J.L., Fulton, R.S., Sudbrak, R., Wen, G.P., Jones, M.C., Hurles, M.E., Andrews, T.D., Scott, C.E., Searle, S., Ramser, J., Whittaker, A., Deadman, R., Carter, N.P., Hunt, S.E., Chen, R., Cree, A., Gunaratne, P., Havlak, P., Hodgson, A., Metzker, M.L., Richards, S., Scott, G., Steffen, D., Sodergren, E., Wheeler, D.A., Worley, K.C., Ainscough, R., Ambrose, K.D., Ansari-Lari, M.A., Aradhya, S., Ashwell, R.I.S., Babbage, A.K., Bagguley, C.L., Ballabio, A., Banerjee, R., Barker, G.E., Barlow, K.F., Barrett, I.P., Bates, K.N., Beare, D.M., Beasley, H., Beasley, O., Beck, A., Bethel, G., Blechschmidt, K., Brady, N., Bray-Allen, S., Bridgeman, A.M., Brown, A.J., Brown, M.J., Bonnin, D., Bruford, E.A., Buhay, C., Burch, P., Burford, D., Burgess, J., Burrill, W., Burton, J., Bye, J.M., Carder, C., Carrel, L., Chako, J., Chapman, J.C., Chavez, D., Chen, E., Chen, G., Chen, Y., Chen, Z.J., Chinault, C., Ciccodicola, A., Clark, S.Y., Clarke, G., Clee, C.M., Clegg, S., Clerc-Blankenburg, K., Clifford, K., Cobley, V., Cole, C.G., Conquer, J.S., Corby, N., Connor, R.E., David, R., Davies, J., Davis, C., Davis, J., Delgado, O., DeShazo, D., Dhami, P., Ding, Y., Dinh, H., Dodsworth, S., Draper, H., Dugan-Rocha, S., Dunham, A., Dunn, M., Durbin, K.J., Dutta, I., Eades, T., Ellwood, M., Emery-Cohen, A., Errington, H., Evans, K.L., Faulkner, L., Francis, F., Frankland, J., Fraser, A.E., Galgoczy, P., Gilbert, J., Gill, R., Glockner, G., Gregory, S.G., Gribble, S., Griffiths, C., Grocock, R., Gu, Y.H., Gwilliam, R., Hamilton, C., Hart, E.A., Hawes, A., Heath, P.D., Heitmann, K., Hennig, S., Hernandez, J., Hinzmann, B., Ho, S., Hoffs, M., Howden, P.J., Huckle, E.J., Hume, J., Hunt, P.J., Hunt, A.R., Isherwood, J., Jacob, L., Johnson, D., Jones, S., de Jong, P.J., Joseph, S.S., Keenan, S., Kelly, S., Kershaw, J.K., Khan, Z., Kioschis, P., Klages, S., Knights, A.J., Kosiura, A., Kovar-Smith, C., Laird, G.K., Langford, C., Lawlor, S., Leversha, M., Lewis, L., Liu, W., Lloyd, C., Lloyd, D.M., Loulseged, H., Loveland, J.E., Lovell, J.D., Lozado, R., Lu, J., Lyne, R., Ma, J., Maheshwari, M., Matthews, L.H., McDowall, J., McLaren, S., McMurray, A., Meidl, P., Meitinger, T., Milne, S., Miner, G., Mistry, S.L., Morgan, M., Morris, S., Muller, I., Mullikin, J.C., Nguyen, N., Nordsiek, G., Nyakatura, G., O'Dell, C.N., Okwuonu, G., Palmer,

S., Pandian, R., Parker, D., Parrish, J., Pasternak, S., Patel, D., Pearce, A.V., Pearson, D.M., Pelan, S.E., Perez, L., Porter, K.M., Ramsey, Y., Reichwald, K., Rhodes, S., Ridler, K.A., Schlessinger, D., Schueler, M.G., Sehra, H.K., Shaw-Smith, C., Shen, H., Sheridan, E.M., Shownkeen, R., Skuce, C.D., Smith, M.L., Sotheran, E.C., Steingruber, H.E., Steward, C.A., Storey, R., Swann, R.M., Swarbreck, D., Tabor, P.E., Taudien, S., Taylor, T., Teague, B., Thomas, K., Thorpe, A., Timms, K., Tracey, A., Trevanion, S., Tromans, A.C., d'Urso, M., Verduzco, D., Villasana, D., Waldron, L., Wall, M., Wang, Q.Y., Warren, J., Warry, G.L., Wei, X.H., West, A., Whitehead, S.L., Whiteley, M.N., Wilkinson, J.E., Willey, D.L., Williams, G., Williams, L., Williamson, A., Williamson, H., Wilming, L., Woodmansey, R.L., Wray, P.W., Yen, J., Zhang, J.K., Zhou, J.L., Zoghbi, H., Zorilla, S., Buck, D., Reinhardt, R., Poustka, A., Rosenthal, A., Lehrach, H., Meindl, A., Minx, P.J., Hillier, L.W., Willard, H.F., Wilson, R.K., Waterston, R.H., Rice, C.M., Vaudin, M., Coulson, A., Nelson, D.L., Weinstock, G., Sulston, J.E., Durbin, R., Hubbard, T., Gibbs, R.A., Beck, S., Rogers, J. and Bentley, D.R. (2005) The DNA sequence of the human X chromosome. *Nature*, **434**, 325-337.

Rotondo, A., Mazzanti, C., Dell'Osso, L., Rucci, P., Sullivan, P., Bouanani, S., Gonnelli, C., Goldman, D. and Cassano, G.B. (2002) Catechol O-methyltransferase, serotonin transporter, and tryptophan hydroxylase gene polymorphisms in bipolar disorder patients with and without comorbid panic disorder. *American Journal of Psychiatry*, **159**, 23-29.

Rousset, F. (2000) Inferences from Spatial Population Genetics. In Balding, D., Bishop, M. and Cannings, C. (eds.), *Handbook of Statistical Genetics*. Wiley and Sons, Ltd.

Royle, N.J., Clarkson, R.E., Wong, Z. and Jeffreys, A.J. (1988) Clustering of Hypervariable Minisatellites in the Proterminal Regions of Human Autosomes. *Genomics*, **3**, 352-360.

Rubinstein, P., Walker, M., Carpenter, C., Carrier, C., Krassner, J., Falk, C. and Ginsberg, F. (1981) Genetics of HLA disease associations: the use of the haplotype relative risk (HRR) and the 'haplo-delta' (Dh) estimates in juvenile diabetes from three racial groups. *Human Immunology*, **3**, 384.

Ruhlen, M. (1994) *On the origin of languages : studies in linguistic taxonomy*. Stanford University Press, Stanford, Calif.

Ruiz-Linares, A., Ortiz-Barrientos, D., Figueroa, M., Mesa, N., Munera, J.G., Bedoya, G., Velez, I.D., Garcia, L.F., Perez-Lezaun, A., Bertranpetit, J., Feldman, M.W. and Goldstein, D.B. (1999) Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6312-6317.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E.S. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832-837.

Sabol, S.Z., Hu, S. and Hamer, D. (1998) A functional polymorphism in the monoamine oxidase A gene promoter. *Human Genetics*, **103**, 273-279.

Sandoval, C., de la Hoz, A. and Yunis, E. (1993) Estructura Genetica de la Poblacio´ n Colombiana. *Rev Fac Med Univ Nac Colombia*, **41**, 3–14.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463-5467.

Sawa, A. and Snyder, S.H. (2002) Schizophrenia: Diverse approaches to a complex disease. *Science*, **296**, 692-695.

Scacchi, R., Gambina, G., Ruggeri, M., Martini, M.C., Ferrari, G., Silvestri, M., Schiavon, R. and Corbo, R.M. (1999) Plasma levels of apolipoprotein E and genetic markers in elderly patients with Alzheimer's disease. *Neuroscience Letters*, **259**, 33-36.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297-302.

Schinka, J.A., Busch, R.M. and Robichaux-Keene, N. (2004) A meta-analysis of the association between the serotonin transporter gene polymorphism (5-HTTLPR) and trait anxiety. *Molecular Psychiatry*, **9**, 197-202.

Schmid, C.W. and Shen, C.K.J. (1985) The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates. In MacIntyre, R.J. (ed.), *Molecular Evolutionary Genetics*. Plenum Press, New York, pp. 323-358.

Schmutz, J., Martin, J., Terry, A., Couronne, O., Grimwood, J., Lowry, S., Gordon, L.A., Scott, D., Xie, G., Huang, W., Hellsten, U., Tran-Gyamfi, M., She, X.W., Prabhakar, S., Aerts, A., Altherr, M., Bajorek, E., Black, S., Branscomb, E., Caoile, C., Challacombe, J.F., Chan, Y.M., Denys, M., Detter, J.C., Escobar, J., Flowers, D., Fotopulos, D., Glavina, T., Gomez, M., Gonzales, E., Goodstein, D., Grigoriev, I., Groza, M., Hammon, N., Hawkins, T., Haydu, L., Israni, S., Jett, J., Kadner, K., Kimball, H., Kobayashi, A., Lopez, F., Lou, Y.N., Martinez, D., Medina, C., Morgan, J., Nandkeshwar, R., Noonan, J.P., Pitluck, S., Pollard, M., Predki, P., Priest, J., Ramirez, L., Retterer, J., Rodriguez, A., Rogers, S., Salamov, A., Salazar, A., Thayer, N., Tice, H., Tsai, M., Ustaszewska, A., Vo, N., Wheeler, J., Wu, K., Yang, J., Dickson, M., Cheng, J.F., Eichler, E.E., Olsen, A., Pennacchio, L.A., Rokhsar, D.S., Richardson, P., Lucas, S.M., Myers, R.M. and Rubin, E.M. (2004) The DNA sequence and comparative analysis of human chromosome 5. *Nature*, **431**, 268-274.

Schneider, J.A., Peto, T.E.A., Boone, R.A., Boyce, A.J. and Clegg, J.B. (2002) Direct measurement of the male recombination fraction in the human beta-globin hot spot. *Human Molecular Genetics*, **11**, 207-215.

Schneider, S. and Excoffier, L. (1999) Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates very among sites: Application to human mitochondrial DNA. *Genetics*, **152**, 1079-1089.

Schneider, S., Roessli, D. and Excoffier, L. (2000) Arelquin ver. 2000: A software for population data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Schurr, T.G. and Sherry, S.T. (2004) Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: Evolutionary and demographic evidence. *American Journal of Human Biology*, **16**, 420-439.

Scott, W.K., Nance, M.A., Watts, R.L., Hubble, J.P., Koller, W.C., Lyons, K., Pahwa, R., Stern, M.B., Colcher, A., Hiner, B.C., Jankovic, J., Ondo, W.G., Allen, F.H., Goetz, C.G., Small, G.W., Masterman, D., Mastaglia, F., Laing, N.G.,

Stajich, J.M., Slotterbeck, B., Booze, M.W., Ribble, R.C., Rampersaud, E., West, S.G., Gibson, R.A., Middleton, L.T., Roses, A.D., Haines, J.L., Scott, B.L., Vance, J.M. and Pericak-Vance, M.A. (2001) Complete genomic screen in parkinson disease - Evidence for multiple genes. *Jama-Journal of the American Medical Association*, **286**, 2239-2244.

Segurado, R., Detera-Wadleigh, S.D., Levinson, D.F., Lewis, C.M., Gill, M., Nurnberger, J.,I., Jr., Craddock, N., DePaulo, J.R., Baron, M., Gershon, E.S., Ekholm, J., Cichon, S., Turecki, G., Claes, S., Kelsoe, J.R., Schofield, P.R., Badenhop, R.F., Morissette, J., Coon, H., Blackwood, D., McInnes, L.A., Foroud, T., Edenberg, H.J., Reich, T., Rice, J.P., Goate, A., McInnis, M.G., McMahon, F.J., Badner, J.A., Goldin, L.R., Bennett, P., Willour, V.L., Zandi P.P., Liu, J., Gilliam, C., Juo, S.H., Berrettini, W.H., Yoshikawa, T., Peltonen, L., Lonnqvist, J., Nothen, M.M., Schumacher, J., Windemuth, C., Rietschel, M., Propping, P., Maier, W., Alda, M., Grof, P., Rouleau, G.A., Del-Favero, J., Van Broeckhoven, C., Mendlewicz, J., Adolfsson, R., Spence, M.A., Luebbert, H., Adams, L.J., Donald, J.A., Mitchell, P.B., Barden, N., Shink, E., Byerley, W., Muir, W., Visscher, P.M., Macgregor, S., Gurling, H., Kalsi, G., McQuillin, A., Escamilla, M.A., Reus, V.I., Leon, P., Freimer, N.B., Ewald, H., Kruse, T.A., Mors, O., Radhakrishna, U., Blouin, J.L., Antonarakis, S.E.and Akarsu, N. (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. American Journal of Human Genetics, **73(1)**, 49-62.

Seielstad, M.T., Minch, E. and Cavalli-Sforza, L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genetics*, **20**, 278-280.

Seltman, H., Roeder, K. and Devlin, B. (2001) Transmission/Disequilibrium test meets measured haplotype analysis: Family-based association analysis guided by evolution of haplotypes. *American Journal of Human Genetics*, **68**, 1250-1263.

Serretti, A., Cusin, C., Rossini, D., Artioli, P., Dotoli, D. and Zanardi, R. (2004) Further evidence of a combined effect of SERTPR and TPH on SSRIs response in mood disorders. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, **129B**, 36-40.

Serretti, A., Zanardi, R., Rossini, D., Cusin, C., Lilli, R. and Smeraldi, E. (2001) Influence of tryptophan hydroxylase and serotonin transporter genes on fluvoxamine antidepressant activity. *Molecular Psychiatry*, **6**, 586-592.

Sham, P. (1997) *Statistics in Human Genetics*. Hodder Arnold.

Sheffield, V.C., Stone, E.M. and Carmi, R. (1998) Use of isolated inbred human populations for identification of disease genes. *Trends in Genetics*, **14**, 391-396.

Shendure, J., Mitra, R.D., Varma, C. and Church, G.M. (2004) Advanced sequencing technologies: Methods and goals. *Nature Reviews Genetics*, **5**, 335-344.

Shifman, S., Bronstein, M., Sternfeld, M., Pisante-Shalom, A., Lev-Lehman, E., Weizman, A., Reznik, I., Spivak, B., Grisaru, N., Karp, L., Schiffer, R., Kotler, M., Strous, R.D., Swartz-Vanetik, M., Knobler, H.Y., Shinar, E., Beckmann, J.S., Yakir, B., Risch, N., Zak, N.B. and Darvasi, A. (2002) A highly significant association between a COMT haplotype and schizophrenia. *American Journal of Human Genetics*, **71**, 1296-1302.

Silva, W.A., Bonatto, S.L., Holanda, A.J., Ribeiro-dos-Santos, A.K., Paixao, B.M., Goldman, G.H., Abe-Sandes, K., Rodriguez-Delfin, L., Barbosa, M., Paco-Larson, M.L., Petzl-Erler, M.L., Valente, V., Santos, S.E.B. and Zago, M.A.

(2002) Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *American Journal of Human Genetics*, **71**, 187-192.

Sklar, P., Gabriel, S.B., McInnis, M.G., Bennett, P., Lim, Y.M., Tsan, G., Schaffner, S., Kirov, G., Jones, I., Owen, M., Craddock, N., DePaulo, J.R. and Lander, E.S. (2002) Family-based association study of 76 candidate genes in bipolar disorder: BDNF is a potential risk locus. *Molecular Psychiatry*, **7**, 579-593.

Slatkin, M. (1994) Linkage Disequilibrium in Growing and Stable-Populations. *Genetics*, **137**, 331-336.

Slatkin, M. and Excoffier, L. (1996) Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. *Heredity*, **76**, 377-383.

Smith, D.J. and Lusis, A.J. (2002) The allelic structure of common disease. *Human Molecular Genetics*, **11**, 2455-2461.

Sobel, E. and Lange, K. (1996) Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58**, 1323-1337.

Sobel, E., Papp, J.C. and Lange, K. (2002) Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics*, **70**, 496-508.

Sobel, E., Sengul, H. and Weeks, D.E. (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Human Heredity*, **52**, 121-131.

Spence, M.A., Flodman, P.L., Sadovnick, A.D., Baileywilson, J.E., Ameli, H. and Remick, R.A. (1995) Bipolar Disorder - Evidence for a Major Locus. *American Journal of Medical Genetics*, **60**, 370-376.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission Test for Linkage Disequilibrium - the Insulin Gene Region and Insulin-Dependent Diabetes-Mellitus (Iddm). *American Journal of Human Genetics*, **52**, 506-516.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T.T., Hjaltason, O., Birgisdottir, B., Jonsson, H., Gudnadottir, V.G., Gudmundsdottir, E., Bjornsson, A., Ingvarsson, B., Ingason, A., Sigfusson, S., Hardardottir, H., Harvey, R.P., Lai, D., Zhou, M.D., Brunner, D., Mutel, V., Gonzalo, A., Lemke, G., Sainz, J., Johannesson, G., Andresson, T., Gudbjartsson, D., Manolescu, A., Frigge, M.L., Gurney, M.E., Kong, A., Gulcher, J.R., Petursson, H. and Stefansson, K. (2002) Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics*, **71**, 877-892.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S.T., Abeysinghe, S., Krawczak, M. and Cooper, D.N. (2003) Human gene mutation database (HGMD (R)): 2003 update. *Human Mutation*, **21**, 577-581.

Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T., Stanley, S.E., Jiang, R.H., Messer, C.J., Chew, A., Han, J.H., Duan, J.C., Carr, J.L., Lee, M.S., Koshy, B., Kumar, A.M., Zhang, G., Newell, W.R., Windemuth, A., Xu, C.B., Kalbfleisch, T.S., Shaner, S.L., Arnold, K., Schulz, V., Drysdale, C.M., Nandabalan, K., Judson, R.S., Ruano, G. and Vovis, G.F. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, **293**, 489-493.

Stine, O.C., Xu, J.F., Koskela, R., McMahon, F.J., Gschwend, M., Friddle, C., Clark, C.D., McInnis, M.G., Simpson, S.G., Breschel, T.S., Vishio, E., Riskin, K.,

Feilotter, H., Chen, E., Shen, S., Folstein, S., Meyers, D.A., Botstein, D., Marr, T.G. and Depaulo, J.R. (1995) Evidence For Linkage of Bipolar Disorder to Chromosome-18 With a Parent-of-Origin Effect. *American Journal of Human Genetics*, **57**, 1384-1394.

Stoneking, M., Fontius, J.J., Clifford, S.L., Soodyall, H., Arcot, S.S., Saha, N., Jenkins, T., Tahir, M.A., Deininger, P.L. and Batzer, M.A. (1997) Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Research*, **7**, 1061-1071.

Strachan, T.a.R., Andrew P. (2003) *Human Molecular Genetics*. BIOS Scientific Publishers Ltd., Oxford, UK.

Straub, R.E., Lehner, T., Luo, Y., Loth, J.E., Shao, W., Sharpe, L., Alexander, J.R., Das, K., Simon, R., Fieve, R.R., Lerer, B., Endicott, J., Ott, J., Gilliam, T.C. and Baron, M. (1994) A Possible Vulnerability Locus For Bipolar Affective-Disorder On Chromosome 21q22.3. *Nature Genetics*, **8**, 291-296.

Subramanian, S., Mishra, R.K. and Singh, L. (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biology*, **4**, art. no.-R13.

Surratt, C.K., Persico, A.M., Yang, X.D., Edgar, S.R., Bird, G.S., Hawkins, A.L., Griffin, C.A., Li, X., Jabs, E.W. and Uhl, G.R. (1993) A Human Synaptic Vesicle Monoamine Transporter Cdna Predicts Posttranslational Modifications, Reveals Chromosome-10 Gene Localization and Identifies Taqi Rflps. *Febs Letters*, **318**, 325-330.

Tabor, H. K., Risch, N. J. and Myers, R. M. (2002) Candidate-gene approaches for studying complex traits: practical considerations. *Nature Reviews Genetics*, **3**, 1-7.

Tajima, F. (1983) Evolutionary Relationship of DNA-Sequences in Finite Populations. *Genetics*, **105**, 437-460.

Tajima, F. (1989) Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, **123**, 585-595.

Takahashi, N., Miner, L.L., Sora, I., Ujike, H., Revay, R.S., Kostic, V., JacksonLewis, V., Przedborski, S. and Uhl, G.R. (1997) VMAT2 knockout mice: Heterozygotes display reduced amphetamine- conditioned reward, enhanced amphetamine locomotion, and enhanced MPTP toxicity. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9938-9943.

Tarazona-Santos, E. and Santos, F.R. (2002) The peopling of the Americas: A second major migration? *American Journal of Human Genetics*, **70**, 1377-1380.

Teare, M.D., Dunning, A.M., Durocher, F., Rennart, G. and Easton, D.F. (2002) Sampling distribution of summary linkage disequilibrium measures. *Annals of Human Genetics*, **66**, 223-233.

Templeton, A.R. (1995) A Cladistic-Analysis of Phenotypic Associations With Haplotypes Inferred From Restriction-Endonuclease Mapping or Dna-Sequencing .5. Analysis of Case-Control Sampling Designs - Alzheimers-Disease and the Apoprotein-E Locus. *Genetics*, **140**, 403-409.

Templeton, A.R. (2002) Out of Africa again and again. *Nature*, **416**, 45-51.

Terwilliger, J. and Ott, J. (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Human Heredity*, **42**, 337-346.

Tienari, P., Wynne, L.C., Moring, J., Lahti, I., Naarala, M., Sorri, A., Wahlberg, K.E., Saarento, O., Seitamaa, M., Kaleva, M. and Laksy, K. (1994) The Finnish Adoptive Family Study of Schizophrenia - Implications for Family Research. *British Journal of Psychiatry*, **164**, 20-26.

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., BonneTamir, B., SantachiaraBenerecetti, A.S., Moral, P., Krings, M., Paabo, S., Watson, E., Risch, N., Jenkins, T. and Kidd, K.K. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, **271**, 1380-1387.

Tishkoff, S.A., Goldman, A., Calafell, F., Speed, W.C., Deinard, A.S., Bonne-Tamir, B., Kidd, J.R., Pakstis, A.J., Jenkins, T. and Kidd, K.K. (1998) A global haplotype analysis of the myotonic dystrophy locus: Implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *American Journal of Human Genetics*, **62**, 1389-1402.

Torroni, A., Neel, J.V., Barrantes, R., Schurr, T.G. and Wallace, D.C. (1994) Mitochondrial-DNA Clock for the Amerinds and Its Implications for Timing Their Entry into North-America. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 1158-1162.

Torroni, A., Schurr, T.G., Cabell, M.F., Brown, M.D., Neel, J.V., Larsen, M., Smith, D.G., Vullo, C.M. and Wallace, D.C. (1993) Asian Affinities and Continental Radiation of the 4 Founding Native-American Mtdnas. *American Journal of Human Genetics*, **53**, 563-590.

Uhl, G.R., Li, S., Takahashi, N., Itokawa, K., Lin, Z.C., Hazama, M. and Sora, I. (2000) The VMAT2 gene in mice and humans: amphetamine responses, locomotion, cardiac arrhythmias, aging, and vulnerability to dopaminergic toxins. *Faseb Journal*, **14**, 2459-2465.

Ullu, E., Murphy, S. and Melli, M. (1982) Human 7sl Rna Consists of a 140 Nucleotide Middle-Repetitive Sequence Inserted in an Alu Sequence. *Cell*, **29**, 195-202.

Ullu, E. and Tschudi, C. (1984) Alu Sequences Are Processed 7sl Rna Genes. *Nature*, **312**, 171-172.

Underhill, P.A., Jin, L., Zemans, R., Oefner, P.J. and Cavalli-Sforza, L.L. (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 196-200.

Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Lahr, M.M., Foley, R.A., Oefner, P.J. and Cavalli-Sforza, L.L. (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Annals of Human Genetics*, **65**, 43-62.

Underhill, P.A., Shen, P.D., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J.R., Mehdi, S.Q., Seielstad, M.T., Wells, R.S., Piazza, A., Davis, R.W., Feldman, M.W., Cavalli-Sforza, L.L. and Oefner, P.J. (2000) Y chromosome sequence variation and the history of human populations. *Nature Genetics*, **26**, 358-361.

Varilo, T. and Peltonen, L. (2004) Isolates and their potential use in complex gene mapping efforts - Commentary. *Current Opinion in Genetics & Development*, **14**, 316-323.

Vawter, M.P., Freed, W.J. and Kleinman, J.E. (2000) Neuropathology of bipolar disorder. *Biological Psychiatry*, **48**, 486-504.

Verheij, C., Bakker, C.E., Degraaff, E., Keulemans, J., Willemsen, R., Verkerk, A., Galjaard, H., Reuser, A.J.J., Hoogeveen, A.T. and Oostra, B.A. (1993) Characterization and Localization of the Fmr-1 Gene-Product Associated with Fragile-X Syndrome. *Nature*, **363**, 722-724.

Vincent, J.B., Masellis, M., Lawrence, J., Choi, V., Gurling, H.M.D., Parikh, S.V. and Kennedy, J.L. (1999) Genetic association analysis of serotonin system genes in bipolar affective disorder. *American Journal of Psychiatry*, **156**, 136-138.

Wahls, W.P., Wallace, L.J. and Moore, P.D. (1990) The Z-DNA Motif D(Tg)30 Promotes Reception of Information During Gene Conversion Events While Stimulating Homologous Recombination in Human-Cells in Culture. *Molecular and Cellular Biology*, **10**, 785-793.

Wain, H.M., Lovering, R.C., Bruford, E.A., Lush, M.J., Wright, M.W., Povey, S. (2002) Guidelines for Human Gene Nomenclature. *Genomics* **79** (4), 464-470.

Wakeley, J., Nielsen, R., Liu-Cordero, S.N. and Ardlie, K. (2001) The discovery of single-nucleotide polymorphisms - and inferences about human demographic history. *American Journal of Human Genetics*, **69**, 1332-1347.

Wall, J.D. and Pritchard, J.K. (2003a) Assessing the performance of the haplotype block model of linkage disequilibrium. *American Journal of Human Genetics*, **73**, 502-515.

Wall, J.D. and Pritchard, J.K. (2003b) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **4**, 587-597.

Walther, D.J. and Bader, M. (2003) A unique central tryptophan hydroxylase isoform. *Biochemical Pharmacology*, **66**, 1673-1680.

Walther, D.J., Peter, J.U., Bashammakh, S., Hortnagl, H., Voits, M., Fink, H. and Bader, M. (2003) Synthesis of serotonin by a second tryptophan hydroxylase isoform. *Science*, **299**, 76-76.

Wang, W.Y.S., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, **6**, 109-118.

Wang, Y.M., Gainetdinov, R.R., Fumagalli, F., Xu, F., Jones, S.R., Bock, C.B., Miller, G.W., Wightman, R.M. and Caron, M.G. (1997) Knockout of the vesicular monoamine transporter 2 gene results in neonatal death and supersensitivity to cocaine and amphetamine. *Neuron*, **19**, 1285-1296.

Watkins, W.S., Ricker, C.E., Bamshad, M.J., Carroll, M.L., Nguyen, S.V., Batzer, M.A., Harpending, H.C., Rogers, A.R. and Jorde, L.B. (2001) Patterns of ancestral human diversity: An analysis of Alu- insertion and restriction-site polymorphisms. *American Journal of Human Genetics*, **68**, 738-752.

Weir, B. (1996) *Genetic Data Analysis*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Weir, B.S. and Cockerham, C.C. (1984) Estimating F-Statistics for the Analysis of Population- Structure. *Evolution*, **38**, 1358-1370.

Weiss, K.M. and Clark, A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics*, **18**, 19-24.

Werrett, D.J. (1997) The National DNA Database. *Forensic Science International*, **88**, 33-42.

White, T.D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G.D., Suwa, G. and Howell, F.C. (2003) Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature*, **423**, 742-747.

Willeit, M., Praschak-Rieder, N., Neumeister, A., Zill, P., Leisch, F., Stastny, J., Hilger, E., Thierry, N., Konstantinidis, A., Winkler, D., Fuchs, K., Sieghart, W., Aschauer, H., Ackenheil, M., Bondy, B. and Kasper, S. (2003) A polymorphism (5-HTTLPR) in the serotonin transporter promoter gene is associated with DSM-IV depression subtypes in seasonal affective disorder. *Molecular Psychiatry*, **8**, 942-946.

Wilson, J.F., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N. and Goldstein, D.B. (2001) Population genetic structure of variable drug response. *Nature Genetics*, **29**, 265-269.

Wood, B. and Collard, M. (1999) Anthropology - The human genus. *Science*, **284**, 65-71.

Wright, A.F., Carothers, A.D. and Pirastu, M. (1999) Population choice in mapping genes for complex diseases. *Nature Genetics*, **23**, 397-404.

Xu, W.M., Liu, L.Z., Mooslehner, K. and Emson, P.C. (1997) Structural organization of the human vesicular monoamine transporter type-2 gene and promoter analysis using the jelly fish green fluorescent protein as a reporter. *Molecular Brain Research*, **45**, 41-49.

Xu, X., Peng, M., Fang, Z. and Xu, X.P. (2000) The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, **24**, 396-399.

Yan, H., Yuan, W.S., Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143-1143.

Yu, N., Chen, F.C., Ota, S., Jorde, L.B., Pamilo, P., Patthy, L., Ramsay, M., Jenkins, T., Shyue, S.K. and Li, W.H. (2002) Larger genetic differences within Africans than between Africans and Eurasians. *Genetics*, **161**, 269-274.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, **35**, 57-64.

Zhao, H., Zhang, S., Merinkangas, K.R., Wildenauer, D.B., Sun, F. and Kidd, K.K. (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *American Journal of Human Genetics*, **67**, 1824.

Zhivotovsky, L.A., Rosenberg, N.A. and Feldman, M.W. (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *American Journal of Human Genetics*, **72**, 1171-1186.

Zhu, X.F., Luke, A., Cooper, R.S., Quertermous, T., Hanis, C., Mosley, T., Gu, C.C., Tang, H., Rao, D.C., Risch, N. and Weder, A. (2005) Admixture mapping for hypertension loci with genome-scan markers. *Nature Genetics*, **37**, 177-181.

Zietkiewicz, E., Yotova, V., Gehl, D., Wambach, T., Arrieta, I., Batzer, M., Cole, D.E.C., Hechtman, P., Kaplan, F., Modiano, D., Moisan, J.P., Michalski, R. and Labuda, D. (2003) Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. *American Journal of Human Genetics*, **73**, 994-1015.

Zilhao, J.a.d.E., F. (1999) The chronology and taxonomy of the earliest Aurignacian and its implications for the understanding of Neanderthal extinction. *Journal of World Prehistory*, **13**, 1-68.

Zondervan, K.T. and Cardon, L.R. (2004) The complex interplay among factors that influence allelic association. *Nature Reviews Genetics*, **5**, 89-100.

Statement of Contribution to Work

Chapter 2

In chapter 2 all labwork was conducted within the laboratory apart from sequencing reactions of target promoter sequence for all study genes, including the plasmid constructs they constituted. Sequence reactions took place at the UCLA Gonda building core facility or at Macrogen Incorporated, South Korea, following preparation of PCR products by myself (*TPH2* and all plasmid constructs) or Dr. Charles Glatt (*SLC6A4*, *SLC18A2*).

Although I did all stages of the labour intensive transfection experiments at least once three people contributed to this work: myself, Dr. Charles Glatt and Miss Maricel Tampicilic. I was mainly responsible for development of plasmid constructs, transfection experiments and detection of luminescence.

Chapter 3

Genotype data for this chapter was collected at other labs as described. All statistical analyses were conducted by myself.

Chapter 4

DNA was extracted and isolated from tissue samples in other labs in collaboration with ours, as described. All labwork for this chapter was conducted by myself including SSCP analysis, fluorescent-based genotyping on an ABI 377 Sequencer and the end-point Taqman® assay. The fluorescent probes for the Taqman® assay were developed at ABI.

APPENDIX

I have attached a paper published as a result of the research undertaken at UCLA characterising the nucleotide variants of a recently identified, neuronal isoform of the *TPH1* gene, *TPH2*.